RODOLPHE DEVILLERS

CONCEPTION D'UN SYSTÈME MULTIDIMENSIONNEL D'INFORMATION SUR LA QUALITÉ DES DONNÉES GÉOSPATIALES

Thèse de doctorat présentée en cotutelle
à la Faculté des études supérieures de l'Université Laval, Québec
dans le cadre du programme de Sciences géomatiques
pour l'obtention du grade de Philosophiae Doctor (Ph.D.)

FACULTÉ DE FORESTERIE ET DE GÉOMATIQUE
UNIVERSITÉ LAVAL
QUÉBEC

et

UNIVERSITÉ DE MARNE-LA-VALLÉE
INSTITUT FRANCILIEN DES GÉOSCIENCES, FRANCE
pour l'obtention du grade de Docteur en Sciences de l'information géographique

Décembre 2004

Résumé court

L'information géographique est maintenant un produit de masse fréquemment manipulé par des utilisateurs non-experts en géomatique qui ont peu ou pas de connaissances de la qualité des données qu'ils utilisent. Ce contexte accroît significativement les risques de mauvaise utilisation des données et ainsi les risques de conséquence néfaste résultant de ces mauvaises utilisations. Cette thèse vise à fournir à des utilisateurs experts ou des experts en qualité une approche leur permettant d'évaluer la qualité des données et ainsi être à même de conseiller des utilisateurs non-experts dans leur utilisation des données. Cette approche se base sur une structuration des données de qualité dans une base de données multidimensionnelle et une communication dynamique et contextuelle utilisant des indicateurs de qualité affichés dans un système SOLAP (*Spatial On-Line Analytical Processing*) combiné à un système d'information géographique.

Abstract

Nowadays Geographic information is a mass-product often manipulated by users without expertise in geomatics and who have little or no knowledge about the quality of the data being manipulated. Such context significantly increases the risks of data misuse and of negative consequences resulting from these misuses. This thesis aims at providing expert-users and data-quality experts with a new approach allowing them to better evaluate spatial data quality in order to advise non-expert users. This approach is based on the management of quality information within a multidimensional database and on the dynamic and contextual exploration of quality information through quality indicators displayed into a SOLAP system (Spatial On-Line Analytical Processing) built on a Geographical Information System (GIS).

Résumé

Les utilisateurs de données géospatiales sont de plus en plus confrontés au problème complexe de l'évaluation de l'adéquation de données à un usager particulier. Étant donné la disponibilité croissante de sources de données, les jeux de données sont plus que jamais hétérogènes et complexes à interpréter. L'information décrivant la qualité des données est disponible tout en étant cependant hétérogène sémantiquement et spatiallement, inaccessible, hermétique, etc. Aussi, elle finit en pratique par être négligée par la plupart des utilisateurs. En fait, une personne doit pouvoir développer une expertise solide pour comprendre correctement les métadonnées et évaluer l'adéquation de jeux de données (ou d'extraits de ces jeux) à des usages spécifiques. Une telle tâche complexe peut impliquer des milliers de métadonnées partiellement corrélées. En conséquence, des experts en qualité des données doivent pouvoir s'appuyer sur des outils pour identifier des problèmes potentiels ainsi que pour synthétiser les informations nécessaires pour formuler leur opinion dans un rapport impliquant leur responsabilité professionnelle.

Afin de supporter de tels experts dans l'évaluation de l'adéquation à l'utilisation (*fitness for use*), cette thèse présente une approche visant à mieux gérer et communiquer l'information sur la qualité des données grâce à un ensemble de concepts relié aux bases de données décisionnelles et aux techniques de visualisation.

Cette approche repose techniquement sur une combinaison des fonctions d'un SIG avec des technologies d'intelligence décisionnelle (principalement le *On-Line Analytical Processing* - OLAP), afin d'adapter l'approche de tableau de bord exécutif pour fournir des indicateurs interactifs et contextuels décrivant la qualité des données géospatiales.

Un prototype nommé MUM (Manuel à l'Usager Multidimensionnel) est présenté afin d'illustrer cette approche, permettant de communiquer l'information sur la qualité des données à différents niveaux de détails.

Avant-Propos

Après avoir lu de nombreux avant-propos lors de ma revue de littérature, c'est à présent à mon tour d'en rédiger un ! Il semble que tout le monde s'accorde sur le fait que faire une thèse est exigeant, tant pour soi que pour ses proches... je confirme !! Beaucoup de personnes soulignent aussi qu'une thèse est constituée de deux composantes contradictoires, l'une étant la solitude que l'on vit dans l'avancée de ce projet personnel, et l'autre étant le nombre immense de personnes qui ont contribué directement ou indirectement à la réflexion ou au contexte de la thèse en général. Ce sont ces personnes que je voudrais remercier dans cet avant-propos (vu leur nombre je ne vais pas toutes les citer mais mon cœur y est).

Je voudrais tout d'abord remercier les deux personnes qui m'ont permis de me rendre jusque là : mes parents, Claude et Françoise. Grâce à leurs coups de pieds dans les fesses lorsque je ne voulais pas travailler (c'est une image, n'appelez pas la DPJ svp) et leur support financier, ils m'ont permis de me rendre jusque là. Je remercie également chaleureusement ma conjointe Alix qui m'a apporté un grand support, surtout dans les derniers mois de rédaction pendant lesquels notre famille s'est agrandie avec l'arrivée de notre petit garçon...

J'ai eu la chance de faire ma thèse sous la direction de deux personnes qui m'ont beaucoup apportées : Yvan Bédard (Québec) et Robert Jeansoulin (France). Merci beaucoup à vous deux ! J'ai eu la chance (que tout le monde n'a pas) d'avoir deux directeurs humains. Vos conseils, votre bonne humeur, combiné à votre rigueur a été une excellente école.

Je remercie aussi énormément Bernard Moulin pour ses excellents conseils, ainsi que la minutie et la célérité de ses évaluations.

Merci également à mes deux prélecteurs, les Pr. Bernard Cervelle en France et David Coleman au Canada. Votre présence sur mon jury de thèse m'honore. Un coup de chapeau spécial à Bernard Cervelle pour l'efficacité impressionnante qu'il a montré du début à la fin de la cotutelle pour dénouer les méandres de l'administration française.

Je remercie aussi le Dr Sami Faïz de l'INSAT (Tunisie) pour les discussions que nous avons eu à propos de mon projet lors de mon passage à Tunis, ainsi que le Prof. Gary

Hunter de l'université de Melbourne (Australie) qui m'a donné ses commentaires pour mon deuxième article.

Un gros merci aussi aux professionnels de recherche situés du côté de chez SIRS, dans le local jaune-Rona du sous-sol du Casault. Ils ont été d'un grand secours pour de nombreuses questions techniques, scientifiques... et sociales ! Merci donc, par ordre de bureau, à Suzie, Marie-Jo, Sonia, Éveline, Patrick et Martin. Et un gros merci aussi à mes deux compagnons de thèse SIRSiens, Jean Brodeur et Marc Gervais pour les diverses discussions sur mon projet et mille et un autres sujets.

Un gros merci et une grosse bise à Carmen Couture qui s'est montrée la secrétaire la plus efficace, disponible et sympathique des trois universités dans lesquelles j'ai étudié. Merci aussi de manière plus générale au personnel administratif du CRG et du département pour leur aide pendant ces années.

Le financement étant un point crucial dans une thèse, je remercie les différents organismes ayant contribués au financement de cette thèse, ainsi que les personnes ayant rédigé les demandes de subvention ! Merci donc à Yvan, Robert et Geoffrey, au réseau GEOIDE, à la fondation de l'Université Laval, au projet européen REVIGIS, au Ministère de la Recherche Science et Technologie du Québec, au consulat de France à Québec et au CRSNG. Sans ce support financier je n'aurais jamais fini ma thèse (ni commencé d'ailleurs...).

Un merci particulier au Centre d'Information Topographique de Sherbrooke, et à leurs représentants, Jean, Sylvain, François, Daniel, qui m'ont donné la chance de faire un stage qui a été très enrichissant. Un gros merci Jean !

Finalement, merci à tous ceux qui par leur présence ont rendu l'environnement de la thèse agréable. À Québec : les étudiants de l'équipe SIRS, du Centre de recherche en géomatique (CRG) et de l'INRS Géoressources. En France et en Europe : les étudiants de l'équipe de Robert Jeansoulin au CMI (Université de Provence) et ceux du projet européen REVIGIS.

A mes parents, Françoise et Claude,
ma conjointe Alix
et mon fils Kerian

Table des matières

Liste des tableaux

Liste des figures

Chapitre 1 : Introduction

1.1 Contexte de la recherche[1]

Les trente dernières années ont vu des changements majeurs dans le domaine des technologies de l'information. Le réseau Internet permet à présent une diffusion rapide et plus facile de données entre organisations ou individus. La croissance du réseau Internet est quasi-exponentielle. Alors qu'on répertoriait environ 100 000 sites Web en 1996, il y en avait près de 10 millions en 2000 et on enregistre près de 50 millions de sites au début de l'année 2004[2]. On observe également une croissance similaire du nombre d'internautes, de serveurs, ainsi que pour la largeur de la bande passante. Au Canada en 2004, 76% des entreprises sont connectées à Internet, celles-ci représentant 97% de l'économie canadienne[3]. Ce développement est entre autres mis à profit pour la vente de produits et services grâce au commerce électronique dont l'expansion est, elle aussi, de type exponentiel.

Cette évolution affecte de la même manière le domaine de l'information géographique. Ainsi, de nombreux sites Web proposent des données géospatiales pouvant être téléchargées ou commandées, gratuitement ou non, en accès public ou restreint (ex.

---

[1] Noter que les références bibliographiques de chaque chapitre se retrouvent à la fin de ces chapitres
[2] http://www.zakon.org/robert/internet/timeline/
[3] http://e-com.ic.gc.ca/

GeoBase[4], GIS Data Depot[5], Alexandria Digital Library[6], Discovery Portal[7], Photocartothèque québécoise[8]).

La diversité des données géospatiales disponibles et leur hétérogénéité (ex. précision, date de dernière mise à jour, couverture spatiale, formats, classes d'objets représentées, coûts) a suscité l'apparition d'outils de catalogage interrogeables sur Internet (ex. Discovery Portal, IDG Géomatique, Alexandria Digital Library). Ces outils nommés géorépertoires (Proulx et Bédard, 1995; Proulx *et al.*, 1997) ou catalogues de données géographiques permettent aux utilisateurs de sélectionner des jeux de données qui les intéressent en fonction de différents critères tels que l'étendue spatiale ou temporelle représentée par les données, les classes d'objet représentées, la date de la dernière mise à jour, etc. (Létourneau *et al.*, 1998; Guptill, 1999).

Ce contexte général a pour conséquence qu'il est à présent relativement aisé pour un internaute de télécharger sur son poste de travail des données géospatiales représentant des phénomènes d'intérêt pour un territoire donné.

Cette révolution numérique a créé un changement de paradigme (REV!GIS, 2001). Auparavant, un jeu de données était généralement produit pour une application donnée et manipulé par des utilisateurs travaillant souvent dans la même organisation qui a produite ces données. Cependant, plus récemment, on assiste à la création de nombreux jeux de données issus de l'intégration de données hétérogènes, rendus accessibles à divers utilisateurs qui peuvent alors les exploiter pour des applications très différentes et non-anticipées.

De plus, tandis que l'utilisation de données géographiques était surtout réservée à des experts qui les manipulaient à l'aide de logiciels complexes et coûteux, l'information géographique est à présent de plus en plus accessible au grand public, puisqu'elle peut être visualisée à l'aide d'outils simples d'utilisation et peu onéreux, voire gratuits (Goodchild, 1995; Agumya et Hunter, 1997; Curry, 1998; Elshaw Thrall et Thrall, 1999). Cette démocratisation de l'information géographique et des outils de consultation et de traitement

---

4 http://www.geobase.ca/
5 http://data.geocomm.com/
6 http://www.alexandria.ucsb.edu/
7 http://geodiscover.cgdi.ca/

a atteint un point tel que, à titre d'exemple, il est maintenant possible d'acheter à peu de frais dans de nombreuses pharmacies et tabagies du Québec des jeux de données géospatiales et leur outil de visualisation afin de planifier ses loisirs (Outils Softmap[9] pour la chasse et pêche, *quad*, randonnée, etc.). L'accroissement des applications géomatiques sur les technologies nomades et les téléphones mobiles devrait encore accroître le phénomène de démocratisation de l'information géographique. Il est donc à présent fréquent que des usagers n'ayant pas d'expertise dans le domaine de l'information géographique aient accès à ce type d'information pour des objectifs professionnels ou privés, souvent à des fins différentes de celles envisagées par le producteur.

1.2 Problématique

1.2.1 Démocratisation des données géospatiales et prise de décision

Étant donné l'augmentation des utilisateurs non-experts dans le domaine de l'information géographique pouvant manipuler ce type de données, ainsi que l'hétérogénéité des sources de données, et donc de leur qualité, l'utilisation de données géospatiales dans des processus de prise de décision n'est pas toujours faite de manière avertie. La probabilité que les usagers considèrent les informations affichées par les systèmes comme exactes est forte, étant donné leur représentation numérique (Chrisman, 1990; Morrison, 1995). Les données numériques donnent ainsi aux utilisateurs une fausse impression d'exactitude, de complétude et de qualité, en raison de leur nature technique et de la grande précision des résultats fournis par les SIG (ex. une mesure de distance faite avec ArcGIS 8.0 est donnée avec six décimales et ce, quelle que soit l'exactitude des données).

Hunter (1999) mentionne que les cartes traditionnelles contenaient généralement dans leurs marges certaines informations quantitatives concernant la précision de celles-ci, telles que des estimations des erreurs de positions horizontale et verticale. Il remarque toutefois que « cette approche, cependant, suppose une connaissance de la part des utilisateurs permettant de savoir jusqu'où les cartes peuvent être crédibles. Malheureusement, dans l'âge numérique, la plupart de ces informations manquent aux résultats des SIG; les nouveaux

[8] http://photocartotheque.mrnfp.gouv.qc.ca
[9] http://www.softmaptech.com

utilisateurs de ces informations sont également souvent inconscients des pièges potentiels pouvant résulter de mauvaises utilisations des données et des technologies associées » (traduction libre) (Hunter, 1999 - p. 633).

Dans la pratique, les cas de mauvaise utilisation de l'information géographique sont fréquemment cités dans la littérature scientifique, les médias et les cas de jurisprudence (Blackmore, 1985; Beard, 1989; Monmonier, 1994; Curry, 1998; Epstein *et al.*, 1998; Hunter, 2001; Gervais, 2004). Les conséquences de mauvaises manipulations sont la plupart du temps minimes. Curry cite comme exemple la mauvaise interprétation faite des cartes utilisant une projection conforme. Il est fréquent que des personnes connaissant peu la cartographie déduisent en voyant ces cartes que, par exemple, la superficie de l'Afrique et du Groenland sont à peu près identiques. Toutefois de nombreux cas ont eu des conséquences plus graves et ont causé des pertes de vies humaines ou des dégâts matériels majeurs, ces cas ayant souvent fini devant des tribunaux (Gervais, 2004).

Étant donné que les données géospatiales sont de plus en plus utilisées dans les processus de prise de décision et dans des domaines de plus en plus variés, les cas de mauvaise utilisation et donc d'accidents et de litiges, ont de fortes chances d'augmenter (Epstein *et al.*, 1998). Hunter (1999) pense même que cette tendance pourrait aller jusqu'à la remise en cause de l'utilisation des systèmes d'information géographique.

Afin de réduire ces risques de mauvaise utilisation, les utilisateurs non-experts devraient pouvoir mieux évaluer l'adéquation de ces données à leur utilisation (*fitness for use*). Toutefois, il est difficile, voire impossible, pour ces utilisateurs non-experts d'évaluer l'adéquation des données, cette évaluation impliquant de nombreuses caractéristiques, documentées à différents niveaux de détails et généralement communiquées dans un langage hermétique pour des non-experts. D'où la nécessité de faire appel à un expert.

1.2.2 Problématique juridique

En complément des problèmes potentiels de mauvaises utilisations résultant de la démocratisation des données, il existe une problématique juridique significative qui suscite un intérêt croissant (Gervais, 2004). Gervais a fait une analyse juridique poussée de différents aspects reliés aux bases de données numériques et à l'information géographique

dans plusieurs pays (ex. Canada, France, Belgique, États-Unis) ainsi qu'à travers l'analyse de 225 causes juridiques. Il a ainsi identifié dans tous les pays étudiés un haut niveau d'incertitude concernant plusieurs aspects juridiques tels que la propriété intellectuelle, les contrats de ventes de données et de services, la responsabilité civile des producteurs d'information géographique. Découlant de ce constat, Gervais identifie un ensemble de tâches que les producteurs de données devraient réaliser pour se conformer à la législation. Parmi ces tâches, les producteurs de données géospatiales doivent fournir aux utilisateurs des informations correctes, complètes et compréhensibles concernant les jeux de données qu'ils fournissent. Ces informations doivent être informatives quant à la qualité des données fournies. Beaucoup de producteurs de données fournissent des informations aux utilisateurs par le biais des métadonnées (c.à.d. données sur les données), celles-ci incluant parfois certaines informations sur la qualité. Toutefois, Gervais identifie plusieurs limitations concernant les métadonnées qui les rendent insuffisantes pour répondre aux obligations légales des producteurs, dont en particulier leur technicité pour des utilisateurs non-experts.

Gervais démontre dans ses travaux l'importance d'avoir recours à l'opinion d'un utilisateur expert ou un expert en qualité qui engagerait sa responsabilité pour évaluer l'adéquation de jeux de données à une utilisation définie (évaluer le *fitness for use*). Ces experts auraient alors besoin d'outils leur présentant les différents aspects de la qualité pour les aider dans cette tâche. Il existe donc un besoin pour des outils permettant de structurer et de communiquer l'information sur la qualité à des utilisateurs experts ou des experts en qualité.

1.3 Hypothèse et objectifs de la recherche

L'hypothèse principale de la thèse est qu'il est possible de fournir aux utilisateurs experts ou aux experts en qualité des indicateurs renseignant sur les différentes caractéristiques de la qualité. Ces indicateurs de qualité peuvent être communiqués de manière contextuelle et à différents niveaux de détails et être intégrés dans un système plus large permettant de supporter les experts dans l'évaluation de l'adéquation des données à une utilisation. La sous-hypothèse est que ce système pourrait être basé sur une combinaison de bases de

données multidimensionnelles, d'outils cartographiques et d'approche du domaine du *Business Intelligence*.

Afin de démontrer cette hypothèse, l'objectif principal de la thèse est de proposer une nouvelle approche pour gérer des données décrivant la qualité des données qu'un usager manipule et les diffuser sous une forme plus compréhensible à des usagers experts ou des experts en qualité de données géospatiales.

Plus précisément, les objectifs spécifiques sont:

- Voir la faisabilité d'utiliser des indicateurs, des tableaux de bord et la technologie SOLAP[10] pour communiquer des informations sur la qualité et identifier les caractéristiques que devrait avoir un outil regroupant ces différentes approches;
- Concevoir un modèle permettant une gestion à différents niveaux de détails des informations relatives à la qualité des données à référence spatiale puis développer, comme preuve de concept, un prototype permettant (1) d'informer l'usager de manière contextuelle sur les différents aspects de la qualité des données géospatiales qu'il manipule et (2) représenter la variabilité spatiale de la qualité des données.

1.4 Méthodologie

Ce projet de recherche a été mené en complémentarité avec celui de Marc Gervais, étudiant au doctorat en Sciences Géomatiques à l'Université Laval ayant terminé en 2004. Marc Gervais a exploré différentes considérations légales reliées aux données numériques géospatiales (ex. responsabilité civile, droits d'auteur), servant en partie de motivation au présent projet. En pratique, les deux projets ont été menés de front à un an d'intervalle, les résultats de Marc Gervais ayant été intégré au fur et à mesure pour orienter la présente thèse. Ce projet, ainsi que celui de Marc Gervais, ont été fait en partie au sein du projet européen REVIGIS[11] (projet IST-1999-14189) portant sur l'utilisation de méthodes de l'intelligence artificielle pour la révision d'information géographique incertaine. Ce projet

[10] Le SOLAP (*Spatial On-Line Analytical Processing*) est une extension spatiale des outils OLAP utilisés dans le domaine du *Business Intelligence*. Cette association permet d'obtenir des outils de support à la prise de décision rapides, permettant à l'utilisateur de naviguer dans les données à différents niveaux de détail et sous différentes formes (ex. carte, tableaux, histogrammes). Ces outils sont présentés en détail dans la section 4.7.1.

[11] http://www.cmi.univ-mrs.fr/REVIGIS/Full/

regroupait des partenaires universitaires provenant de six pays ainsi qu'un partenaire industriel (SOMEI/Marseille). La contrepartie québécoise du projet était le projet du Ministère de la Recherche Science et Technologie « développement de technologies de fusion de données géospatiales ». Cette thèse ayant été effectuée en cotutelle France/Québec, trois trimestres ont été passés en France, au Centre de Mathématiques et d'Informatique de Marseille, et ont ainsi permis à travers des réunions et discussions, de plus interagir avec les autres partenaires du projet REVIGIS . Ce contexte de recherche a ainsi permis d'explorer les idées avec un grand nombre et une grande diversité d'intervenants universitaires s'étalant de la géographie à l'intelligence artificielle ainsi que gouvernementaux et industriels.

La méthodologie générale suivie dans cette thèse est présentée sur la Figure 1.

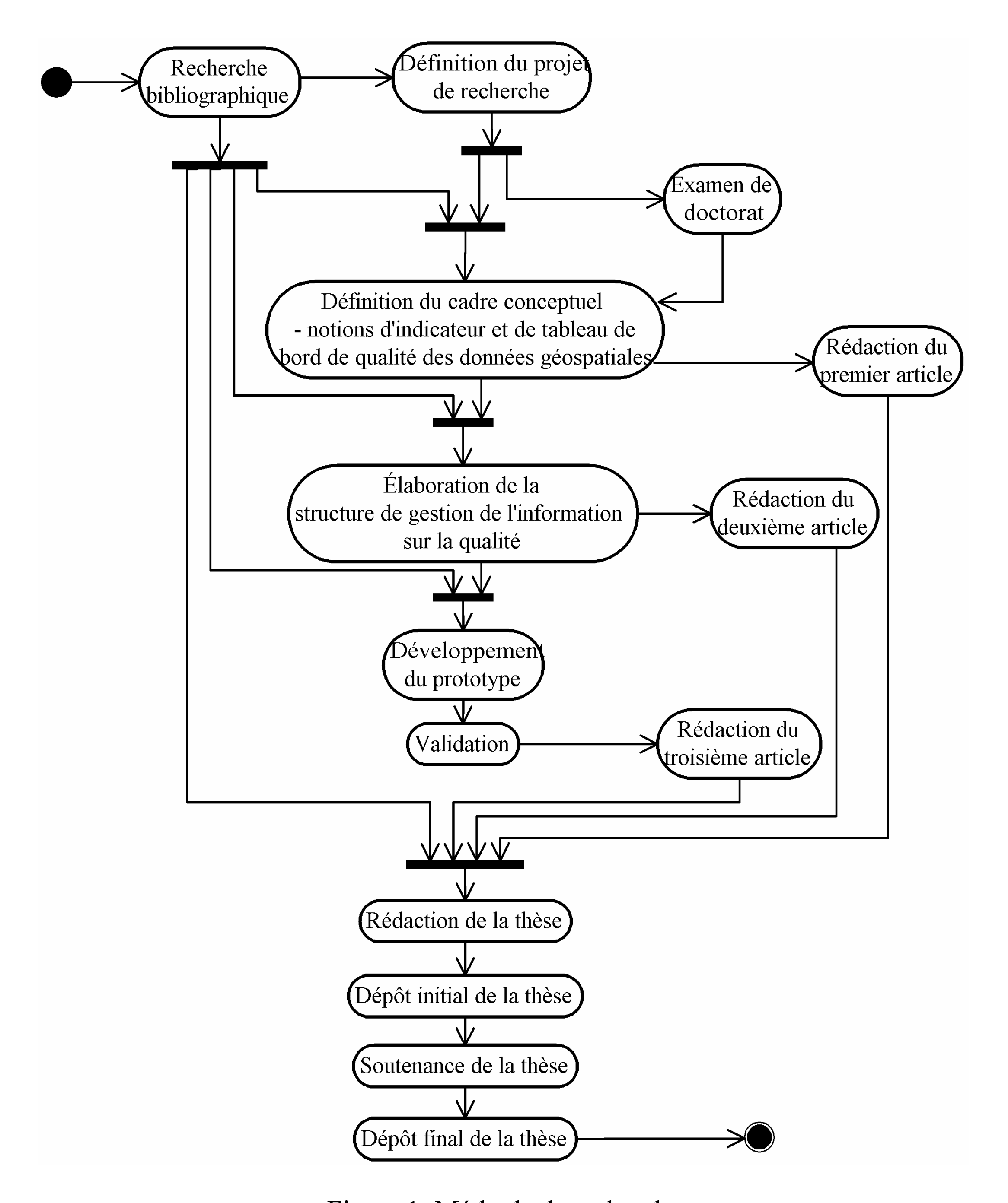


Figure 1: Méthode de recherche

Une recherche bibliographique a été effectuée de manière intensive au début du projet afin de mieux cerner le contexte de la recherche, d'identifier les hypothèses et les objectifs de manière plus précise et de se positionner de manière générale par rapport aux travaux précédemment publiés. Par la suite, tout au long du projet, des recherches bibliographiques

ont été effectuées pour chaque nouveau thème abordé. Une veille bibliographique a également été faite jusqu'à la fin du projet afin d'identifier les nouvelles publications pertinentes pour le projet. La recherche bibliographique effectuée a couvert différents sujets tels que: la qualité des données géospatiales (acquisition, évaluation, gestion, communication, visualisation, utilisation), les métadonnées (normalisation, production et utilisation), le domaine de la prise de décision, de la gestion du risque, des bases de données traditionnelles (relationnelles) et analytiques (multidimensionnelles), les techniques du *Business Intelligence* ou adaptées de ce domaine (ex. OLAP, SOLAP, tableaux de bord de gestion, entrepôts de données) ainsi que des aspects légaux reliés à l'information géographique.

Basé sur cette revue de littérature considérant environ 250 articles et livres, le projet de recherche a été défini, détaillant le contexte, les questions, hypothèses et objectifs de la recherche, une synthèse de la littérature, puis une méthodologie incluant les étapes et un échéancier. Ce projet a par la suite été défendu devant un comité à l'oral et à l'écrit lors d'un examen de doctorat.

La deuxième étape a porté sur l'élaboration des notions d'indicateurs et de tableau de bord de qualité pour les données géospatiales. Cette étape a eu pour objectif de voir dans quelle mesure il est possible d'utiliser des indicateurs, approche couramment utilisée dans le domaine de la gestion, comme outil de communication des informations sur la qualité des données géospatiales. Basé sur la revue de littérature faite dans le domaine des indicateurs et de la prise de décision, un cadre théorique a été développé pour adapter cette approche au domaine de la géomatique. Les caractéristiques que devrait avoir le système ont été identifiées. Une maquette visuelle a été développée à cette étape afin de préciser les caractéristiques qu'aurait une interface cartographique incluant des indicateurs de qualité. La maquette a été présentée à différents intervenants du domaine de la géomatique provenant des milieux universitaires, gouvernementaux et industriels.

La troisième étape a porté sur la définition d'un modèle permettant de gérer l'information sur la qualité. Ce modèle permet à la fois de gérer l'information sur la qualité à différents niveaux de détails, mais intègre également une hiérarchisation des indicateurs de qualité. Un modèle multidimensionnel a été proposé, permettant ainsi de bénéficier des opérateurs

de navigation fournis par les systèmes de type SOLAP ainsi que des courts temps de réponse de ces systèmes.

La quatrième étape a consisté en un prototypage informatique permettant de valider les concepts développés dans les deux étapes précédentes. Basé sur les résultats de l'analyse et de la conception, une partie de l'implémentation du prototype (i.e. chargement des données et programmation) a été effectuée dans le cadre d'un stage de 3 mois d'un étudiant au baccalauréat de 4ème année en Sciences Géomatiques, Mathieu Lachapelle (dirigé par Yvan Bédard et encadré par Rodolphe Devillers). Le prototype a été développé en Visual Basic, combinant différentes technologies : SIG (Intergraph GeoMedia), base de données relationnelle (Microsoft Access), base de données multidimensionnelles (Microsoft SQL Serveur) et un client OLAP (Proclarity). Le feuillet cartographique 021e05 de la Base Nationale de Données Topographiques du Canada[12] (BNDT) (échelle 1 :50 000), a été utilisé pour le prototype. Ce jeu de données a été sélectionné pour deux raisons principales : (1) c'est un produit qui possède des métadonnées mieux documentées que la moyenne et allant jusqu'à une description des primitives géométriques et (2) le Centre d'Information Topographique de Sherbrooke (CIT-S), organisme produisant ces données, était partenaire du projet européen REVIGIS dans lequel s'insérait partiellement ce projet. Le CIT-S a de plus fourni gratuitement leurs données. Ce feuillet représente le centre de la ville de Sherbrooke (Québec) et a la particularité d'inclure des zones urbaines et plus rurales pouvant avoir des qualités différentes. Le fait que plusieurs municipalités récemment fusionnées soient présentes sur le feuillet a également un intérêt au regard de l'hétérogénéité de la qualité des données. Un sous-ensemble géographique et thématique du feuillet a été fait pour les fins d'expérimentation. Parmi plus de 110 classes d'objets disponibles[13], les classes d'objets représentant les routes (*roadl, li_roal*), les cours d'eau (*watercl, waterbd*) et les bâtiments (*buildid, buildip, builtud*) ont été utilisées.

Une validation de l'approche et du prototype a été faite à différents stades du projet, tant sur le plan scientifique que sur le plan de l'utilité de l'approche pour différents types d'utilisateurs. Sur le plan scientifique, l'approche développée dans ce projet a fait l'objet de 13 communications scientifiques dans des revues, conférences nationales et internationales,

---

[12] http://www.cits.rncan.gc.ca/
[13] http://scar.cits.rncan.gc.ca/bndt/bndt.htm

etc., présentant différents aspects du projet à différents stades de réflexion. Cela a permis de discuter et de valider le contenu scientifique avec des experts en qualité, en bases de données, et en géomatique en général. Le projet a donné lieu à différentes présentations et discussions lors des rencontres du projet REVIGIS pendant les quatre années de la thèse, et s'est également partiellement inscrit dans les projets GEOIDE[14] SOC#1 et DEC#2 au début de la thèse. Sur le plan de l'utilité de l'approche, les concepts ainsi que le système développé ont été présentés à différents types d'utilisateurs, incluant des experts et non-experts en géomatique, des thématiciens, des représentants d'agences gouvernementales (ex. Santé Canada, Défense Canada, Géomatique Canada, Ministère des Ressources Naturelles du Québec), d'industries (ex. Kheops Technologies, Hydro-Québec, SOMEI/Société des eaux de Marseille, Swiftsure Spatial Systems), etc. Un stage de deux mois a été effectué en 2002 au sein du Centre d'Information Topographique de Sherbrooke, sous la direction du Dr. Jean Brodeur, afin de mieux appréhender les considérations reliées à la production de données et de métadonnées numériques ainsi qu'à l'utilisation faite des métadonnées par leurs clients. Ces discussions ont permis de souligner que le problème de communication de la qualité est une préoccupation croissante, commune aux différents domaines utilisant des données géospatiales, et que la solution proposée dans cette thèse est d'intérêt pour différents types d'utilisateurs.

Finalement, la dernière étape a consisté à intégrer les articles écrits dans la présente thèse, rédiger une revue de littérature plus complète et cohérente (chapitre 2) et des chapitres d'introduction et de conclusion.

1.5 Présentation de la thèse

Les résultats de la thèse ont été communiqués à travers trois publications principales, soumises à des revues scientifiques à comité de lecture dans le domaine de la géomatique. Ces trois articles constituent le cœur de la thèse et sont présentés dans les chapitres 3, 4 et 5 de ce document. Des modifications mineures ont été apportées aux articles afin de mieux les intégrer dans le format de la thèse. Toutefois, le texte des articles n'a pas été significativement modifié par rapport aux versions soumises ou acceptées. Étant donné que les trois articles portent sur le même projet, il existe parfois une certaine redondance entre

[14] Réseau Canadien des Centres d'Excellence en géomatique (http://www.geoide.ulaval.ca)

les articles, celle-ci faisant parfois suite à la demande des évaluateurs des revues, désirant connaître les différentes parties du projet. Toutefois, les articles ayant été écrit à différentes périodes de la thèse, le contenu qui peut sembler redondant est écrit avec différents niveaux de maturation de la réflexion.

Le chapitre 2 présente une revue de littérature plus complète et cohérente que celles présentées dans les articles, permettant ainsi d'introduire divers travaux issus de la littérature pertinents à l'élaboration de la réflexion présentée dans cette thèse. Le chapitre 3 présente l'approche par indicateurs et tableau de bord comme outil de communication de l'information sur la qualité. Le chapitre 4 traite de la gestion à différents niveaux de détails de l'information décrivant la qualité des données. Par la suite, le chapitre 5 présente un prototype, développé dans le cadre de ce projet, visant à implanter et tester les approches décrites dans les chapitres 3 et 4. Finalement, le chapitre 6 conclut la thèse, discute des résultats, identifie les limites de la recherche et ouvre sur de nouvelles perspectives de recherche.

1.6 Références

Agumya A., Hunter G.J., "Determining fitness for use of geographic information", *ITC Journal*, vol. 2, nº 1, 1997, p. 109-113.

Beard K., "Use error: the neglected error component", *Proceedings of AUTO-CARTO 9*, Baltimore, Maryland, Mars 1989, p. 808-817.

Blackmore M., "High or Low Resolution? Conflicts of Accuracy, Cost, Quality and Application in Computer Mapping", *Computers & Geosciences*, vol. 11, nº 2, 1985, p. 345-348.

Chrisman N.R., "The error component in spatial data". *Geographic Information Systems: Principles and Applications* (D.J. Maguire, M.F. Goodchild et D.W. Rhind, Eds), Wiley, London, p. 165-174, 1990.

Curry M.R., *Digital Places: Living with Geographic Information Technologies*, London & New-York, Routeledge, 1998.

Elshaw Thrall S., Thrall G.I., "Desktop GIS software". *Geographical Information Systems* (P. A. Longley, M. F. Goodchild, D.J. Maguire et D.W. Rhind, Eds), John Wiley & Sons, New-York, p. 331-345, 1999.

Epstein E.F. Hunter G.J., Agumya A., "Liability insurance and the use of geographical information", *International Journal of Geographical Information Science*, vol. 12, nº 3, 1998, p. 203-214.

Gervais M., *Pertinence d'un manuel d'instructions au sein d'une stratégie de gestion du risque juridique découlant de la fourniture de données géographiques numériques*, Thèse de doctorat, Département des Sciences Géomatiques, Université Laval, Québec, 2004.

Goodchild M.F., "Sharing Imperfect Data". *Sharing Geographic Information* (H.J. Onsrud et G. Rushton, Eds), Rutgers University Press, New Brunswick, NJ, p. 413-425, 1995.

Guptill S.C., "Metadata and data catalogues". *Geographical Information Systems* (P.A. Longley, M.F. Goodchild, D.J. Maguire et D.W. Rhind, Eds), John Wiley & Sons, Inc., p. 677-692, 1999.

Hunter G.J., "Managing uncertainty in GIS". *Geographical Information Systems* (P.A. Longley, M.F. Goodchild, D.J. Maguire et D.W. Rhind, Eds), John Wiley & Sons, Inc., p. 633-641, 1999.

Hunter G.J., "Spatial Data Quality Revisited", *Proceedings of GeoInfo 2001*, Rio de Janeiro, Brésil, 4-5 octobre 2001, p. 1-7.

Létourneau F. Bédard Y., Moulin B., "Perspectives d'utilisation du concept d'entrepôt de données pour les géorépertoires dans internet", *Geomatica*, vol. 52, nº 2, 1998, p. 145-163.

Monmonier M., "A Case Study in the Misuse of GIS: Siting a Low-Level Radioactive Waste Disposal Facility in New-York State", *Proceedings of Conference on Law and Information Policy for Spatial Databases* (H. Onsrud, Ed.), Tempe (AZ) USA, 1994, p. 293-303.

Morrison J.L., "Spatial data quality". *Elements of spatial data quality* (S.C. Guptill et J.L. Morrison, Eds), Elsevier Science inc., New York, 1995.

Proulx M.J., Bédard Y., "Le géorépertoire, un outil de gestion cartographique", *Arpenteur-Géomètre, Revue de l'Ordre des Arpenteurs-Géomètres du Québec*, vol. 21, nº 5, 1995, p. 21-24.

Proulx M.J. Bédard Y. Létourneau F., Martel C., "Catalogage des données spatiales sur le world wide web: concepts, analyses des sites et présentation du géorépertoire personnalisable GEOREP", *Revue Internationale de Géomatique*, vol. 7, nº 1, 1997, p. 7-32.

REV!GIS, 2001. Uncertain Knowledge Maintenance and Revision in Geographic Information Systems, Projet européen IST-1999-14189, http://www.lsis.org/REVIGIS.

Chapitre 2 : Revue de littérature

La recherche abordée par cette thèse nécessite la compréhension de différents concepts reliés, entre autres, aux domaines des systèmes d'information géographiques et des bases de données. Ce chapitre présente une synthèse de la littérature portant sur différents concepts jugés pertinents pour cette thèse.

Nous présentons dans un premier temps la place des données géospatiales et des SIG dans les processus de prise de décision, mettant l'accent sur l'importance des imperfections reliées aux données géospatiales. Dans un deuxième temps, nous présentons le concept de qualité, central dans cette thèse, ainsi que la terminologie gravitant autour de ce terme. Nous examinons ici le concept de qualité de manière générale puis nous nous intéressons plus spécifiquement à la qualité des données géospatiales. Nous présentons ensuite les différentes étapes menant à la communication de l'information sur la qualité, soit l'évaluation de la qualité de données géospatiales, la gestion de ces informations décrivant la qualité, puis les approches permettant de communiquer ces informations. Finalement, une synthèse générale des constats faits dans ce chapitre est présentée afin d'appuyer l'approche suivie dans cette thèse.

2.1 Systèmes d'information géographique et processus de prise de décision

Les systèmes d'information géographiques sont de plus en plus utilisés pour supporter des processus de prise de décision. Cette section montre (1) que de l'incertitude est inhérente aux données géospatiales, (2) que cette incertitude devrait être prise en compte lors de l'utilisation des données et (3) que la communication des données géospatiales, et aussi de l'incertitude, passent par l'utilisation d'un langage plus proche de celui utilisé par les utilisateurs des données.

2.1.1 Information géographique, abstraction et sources d'erreur

Les données géospatiales sont des représentations de phénomènes du monde réel selon des points de vue particulier. Ainsi, pour une étendue spatiale donnée, un plan cadastral pourra représenter le territoire sous la forme de parcelles, tandis qu'une carte topographique représentera ce même territoire sous la forme de bâtiments, rivières, routes, courbes de niveau, etc. Une autre carte topographique pourra également représenter le même territoire à une échelle plus petite, simplifiant certains détails du territoire considérés comme moins utiles sur cette carte (c.à.d. processus de généralisation cartographique). Ainsi, chaque représentation cartographique de l'espace résulte d'abstractions permettant de représenter le territoire de manière simplifiée, suivant un but défini. Ainsi, du fait des processus d'abstraction et de simplification effectués, toutes les cartes papier ou numériques sont à différents niveaux inexactes, incomplètes et inactuelles. La Figure 2 illustre la représentation d'un même phénomène dans un SIG (c.à.d. les routes) pour une même étendue spatiale, mais à différentes échelles allant de 1 :1000 à 1 :250 000. En plus de la différence dans la position des routes, on observe une différence dans le type de représentation, les routes étant représentées au 1 :1000 par l'espace situé entre deux lignes (limites de la route) et par une ligne représentant le centre de la route pour les échelles plus petites.

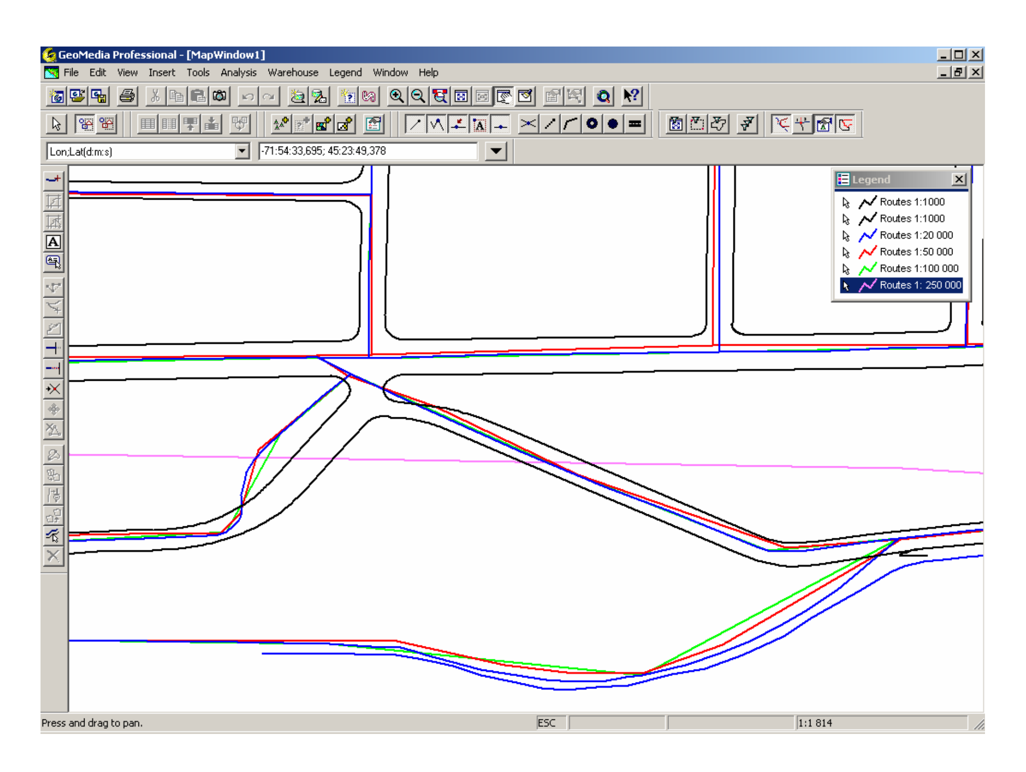


Figure 2 : Routes provenant de jeux de données gouvernementaux et municipaux allant de l'échelle 1 :1000 à 1 :250 000

Ainsi, le statisticien Box (1976) dit que « tous les modèles sont faux, mais certains sont utiles ». De façon similaire, Bédard (1986) dit que les modèles ne sont pas vrais ou faux, mais utiles ou inutiles. Longley *et al.* (2001) mentionnent aussi qu'« il est impossible de produire une représentation parfaite du monde et donc que l'incertitude associée à cette représentation est inévitable ». Eco (2000) présente certaines limites de la cartographie dans son texte « carte de l'empire » dans lequel il montre avec humour les difficultés, et l'absurdité, de produire une représentation à l'échelle 1:1 de la réalité. Bien qu'une carte à cette échelle serait relativement exacte, elle n'aurait que peu d'utilité étant donné que l'un des objectifs initial des cartes et de communiquer une représentation simplifiée (c.à.d. un modèle) de la réalité.

Le processus d'abstraction est donc une première source de différence entre des données produites (selon un certain processus d'abstraction) et des données désirées par l'utilisateur pour une application donnée (Bédard, 1987). Une seconde source de différence est causée par des erreurs qui peuvent affecter les données tout au long de leur processus de production. Les sources d'erreur des données géospatiales sont souvent classifiées en deux

types: les erreurs d'acquisition et les erreurs de traitement (Beard, 1989), ces deux classes étant ensuite souvent divisées en sous-classes. L'introduction et la propagation d'erreurs dans les données sont par exemple reliées aux procédures de collecte des données (ex. précision des instruments, erreurs de calibrage, erreurs de manipulation) ou à leur transformation en des données utilisables (ex. numérisation, vectorisation, généralisation, interpolation, conversion de formats).

2.1.2 Incertitude et prise de décision

Les données à référence spatiale sont de plus en plus utilisées comme support à la prise de décision dans un nombre croissant de domaines d'applications et à différents niveaux organisationnels (c.à.d. opérationnel, tactique et stratégique) (Longley *et al.*, 1999). Toutefois, ces données contiennent toujours un certain niveau d'incertitude, les rendant utiles dans certains contextes et moins dans d'autres. Ainsi, il existe des risques significatifs à utiliser des données non-adéquates dans certains processus de prise de décision.

Goodchild (1995) suggère que les recherches actuelles ne doivent pas uniquement s'intéresser à la description de la qualité des données et à leur transfert aux utilisateurs mais également à la nature de l'impact qu'ont les informations sur la qualité des données sur les processus de décision que les SIG doivent supporter. Il affirme que personne ne peut désirer utiliser des données dans lesquelles il n'a pas confiance ou avec des précisions qu'il ne peut pas comprendre. Goodchild décrit alors les SIG comme étant leur propre ennemi: en invitant les personnes à trouver de nouvelles utilisations pour les données, on les invite à être irresponsables dans leur utilisation. Dans le même sens, Beard (1989) souligne l'importance des problèmes d'utilisation en enrichissant la typologie des erreurs, ajoutant aux erreurs d'acquisition (*source errors*) et de traitement (*process errors*) les erreurs d'utilisation (*use errors*), ce type d'erreur étant rencontré de plus en plus souvent avec la démocratisation des données géospatiales (Epstein *et al.*, 1998). Ces erreurs peuvent conduire à des décisions prises dans un climat d'incertitude.

L'incertitude peut être située à différents niveaux, les différents types d'incertitude étant souvent présents dans un même jeu de données. Fisher (1999) mentionne les problèmes de définition (1) des classes d'objet observées et (2) des objets individuels composant cette

classe, Taylor (1982) identifiant ce problème comme le « problème de définition ». Bédard (1986) classifie l'incertitude en quatre catégories:

- *Conceptuelle* (1er ordre): réfère au flou lors de l'identification d'une réalité observée;
- *Descriptive* (2ème ordre): réfère au manque de précision quant aux valeurs des attributs d'une réalité observée;
- *De localisation* (3ème ordre): réfère au manque de précision dans la localisation dans l'espace et le temps d'une réalité observée;
- *Méta-incertitude* (4ème ordre): réfère au niveau auquel les incertitudes précédentes sont connues.

Lorsqu'un utilisateur fait face à des incertitudes lors d'une prise de décision et qu'il est conscient du type d'incertitude et de son ampleur, il est alors en mesure de choisir entre (1) ne rien faire, (2) essayer de diminuer le niveau d'incertitude ou (3) prendre la décision en acceptant les conséquences possibles, « absorbant » ainsi cette incertitude (Bédard, 1986; Hunter, 1999). Epstein *et al.* (1998) suggèrent que l'incertitude peut être diminuée lorsque (1) on acquiert plus d'information et/ou (2) on améliore la qualité de l'information disponible. L'incertitude résiduelle pouvant être absorbée correspond alors au niveau de risque relié à l'utilisation de l'information (Bédard, 1986; Epstein *et al.*, 1998). Hunter (1999) présente une stratégie globale permettant de gérer l'incertitude dans les SIG intégrant les concepts d'absorption et réduction d'incertitude (cf. Figure 3). Dans cette démarche, une comparaison est faite entre les caractéristiques des données et les besoins des utilisateurs (qualité nécessaire).

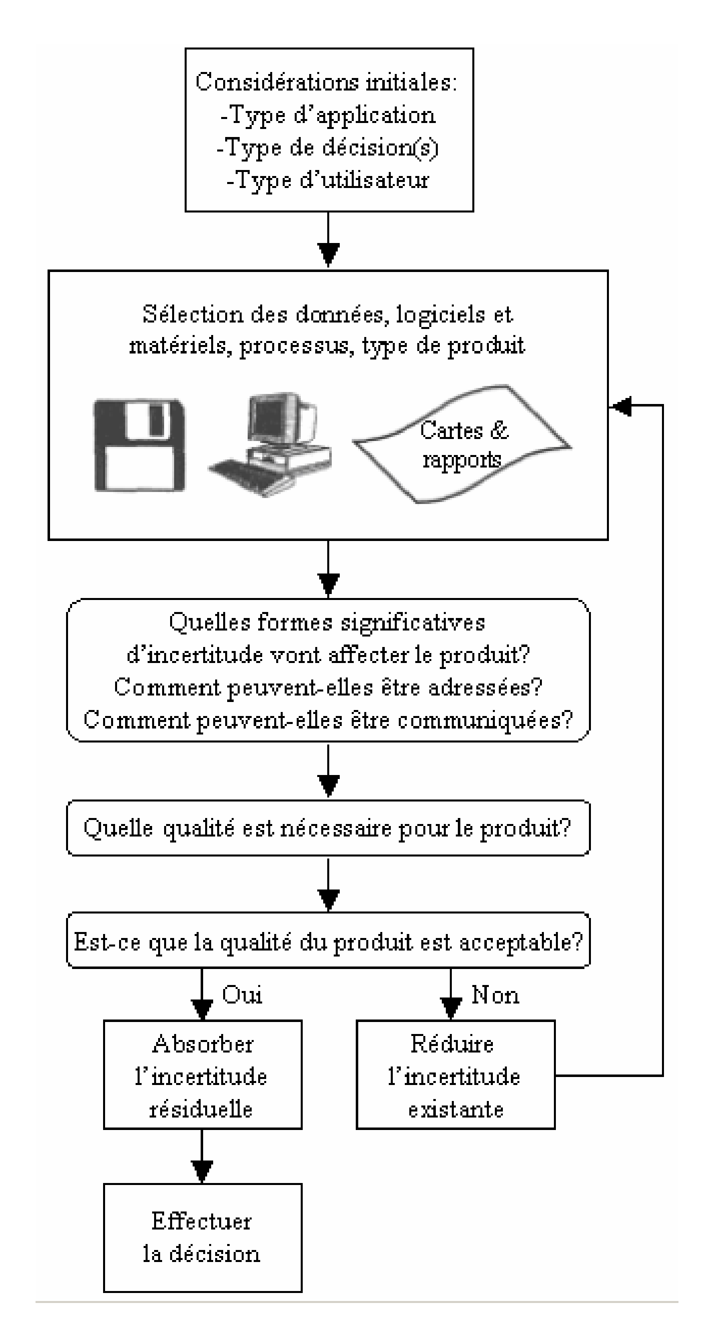


Figure 3: Stratégie de gestion de l'incertitude dans les SIG (traduit de Hunter, 1999)

Certains auteurs proposent des méthodes quantitatives permettant de déterminer l'adéquation entre les caractéristiques des jeux de données et les besoins des utilisateurs (De Bruin *et al.*, 2001). Toutefois, Agumya et Hunter (1997) affirment que la définition de l'adéquation à l'utilisation d'un jeu de données géospatiales dans des applications contextuelles reste le fardeau de l'usager et que la société est pour l'instant mal préparée

pour cette tâche. Ils mentionnent également que cette situation est aggravée par l'absence de modèles et d'outils pouvant aider les usagers dans cette tâche (Agumya et Hunter, 1997). Les auteurs proposent une approche originale pour le domaine, déterminant le niveau acceptable d'incertitude en analysant les risques potentiels pouvant être associés à une prise de décision basée sur ces données. Le risque est ici défini par la probabilité qu'un événement adverse soit la conséquence d'une décision, multiplié par le coût de cet événement. Pour une meilleure compréhension du concept de risque, cette définition peut être complétée par celle utilisée dans le domaine des risques naturels (Manche, 2000), le risque étant l'intersection entre aléas (ex. avalanches, crues, glissement de terrain) et vulnérabilité (ex. zones d'habitation, routes). Ainsi, des avalanches se produisant dans une zone non fréquentée par l'homme ne constituent pas un risque. Cette définition du risque dans un contexte environnemental peut facilement être adaptée aux risques de mauvaise utilisation de l'information géographique, le risque existant à l'intersection des opérations faites avec le SIG et des données de qualité variables. Agumya et Hunter (1997) définissent un processus devant aider à déterminer l'adéquation de jeux de données à un usage spécifique, soit: (1) modélisation, (2) propagation, (3) communication, (4) adéquation à l'utilisation (*fitness for use*) et (5) réduction de l'incertitude.

2.1.3 SIG : un processus de communication

En tant qu'outils, les systèmes d'information géographiques ont pour principaux objectifs de gérer des informations à référence spatiale, de les traiter puis de les communiquer à l'aide par exemple de listes, tableaux ou cartes thématiques. De manière plus générale, Bédard (1987) décrit les systèmes d'information géographiques (SIG) comme étant des processus de communication complexes entre les producteurs et les utilisateurs de données. La communication forme à elle seule un vaste de domaine de recherche duquel la présente recherche s'inspire, beaucoup de modèles de communication ayant été développés (Willett, 1992). Le terme communication peut être défini comme « reproduire en un point un message émis en un autre point, de manière exacte ou approximative » (traduction libre) (Shannon, 1948). Afin de prendre une décision, un individu doit recevoir des signaux du monde réel (observations), interpréter ces signaux puis procéder à une abstraction afin de se créer un modèle cognitif. Un des aspects importants pour un processus de communication

est que les émetteurs et récepteurs (pouvant être des individus ou des machines) doivent partager des connaissances communes (Bédard, 1986; Martinet et Marti, 2001; Brodeur *et al.*, 2003). Plus cette connaissance commune est grande, plus petite sera la distorsion du message entre la source et la cible (Schramm, 1971). En pratique, cette communication n'est jamais parfaite, étant donné les différences existant entre émetteur et récepteur. Dans ce sens, Martinet et Marti encouragent l'utilisation d'un langage proche de celui du récepteur afin de faciliter la transmission des messages dans une entreprise.

2.2 Qualité des données

Cette thèse porte sur la qualité des données géospatiales, et plus spécifiquement sur la gestion et la communication des informations sur la qualité des données. Toutefois, l'utilisation du terme qualité dans la littérature et le langage courant présente beaucoup de variations et est souvent fait de manière incorrecte. Cette section vise à clarifier et définir les différents concepts reliés à la qualité. Nous présentons dans un premier temps certains termes gravitant autour du concept de qualité. Nous présentons ensuite le concept de qualité de manière globale, puis de façon plus spécifique le concept de qualité pour des données géospatiales.

2.2.1 Terminologie de l'incertitude et de l'ignorance

Beaucoup de termes gravitant autour du concept de qualité se retrouvent dans la littérature scientifique (ex. incertitude, erreur, précision, exactitude, vague, flou), ces termes étant souvent employés de manière inexacte. Pour cette raison, plusieurs auteurs (voir par exemple Fisher, 1999 ou Smithson, 1989) ont proposé des définitions de ces termes et les ont mis en relation dans des taxonomies. Les définitions de ces termes sont variées est donnent lieu à de riches discussions dans la communauté. Sans vouloir entrer dans des discussions philosophiques, cette section vise à clarifier l'utilisation qui va être faite de certains termes dans cette thèse. Par exemple, le terme « qualité », allant être décrit en détail dans la section suivante, est la plupart du temps employé ailleurs dans la littérature dans le sens d'imprécision, incertitude, erreur, etc. Des données de qualité sont ainsi souvent uniquement associées à des données ayant une grande précision spatiale. Toutefois, le concept de qualité est bien plus large que la seule notion de précision.

Smithson (1989) propose une taxonomie de l'ignorance. L'ignorance y est vue comme étant multiple et ayant différents niveaux. Elle constitue le concept le plus élevé de sa taxonomie. Elle est, au même titre que la connaissance, une construction sociale, variant selon le point de vue qu'ont d'autres acteurs ayant eux-mêmes une certaine connaissance. Smithson sépare l'ignorance en deux types, soit l'erreur et l'inapplicabilité (*irrelevance*). Le concept d'incertitude est dans cette classification un type particulier d'incomplétude, étant lui-même un type d'erreur (cf. Figure 4). Le terme « incertitude » est souvent employé dans la littérature comme équivalent au concept d'erreur de Smithson, regroupant également l'inexactitude.

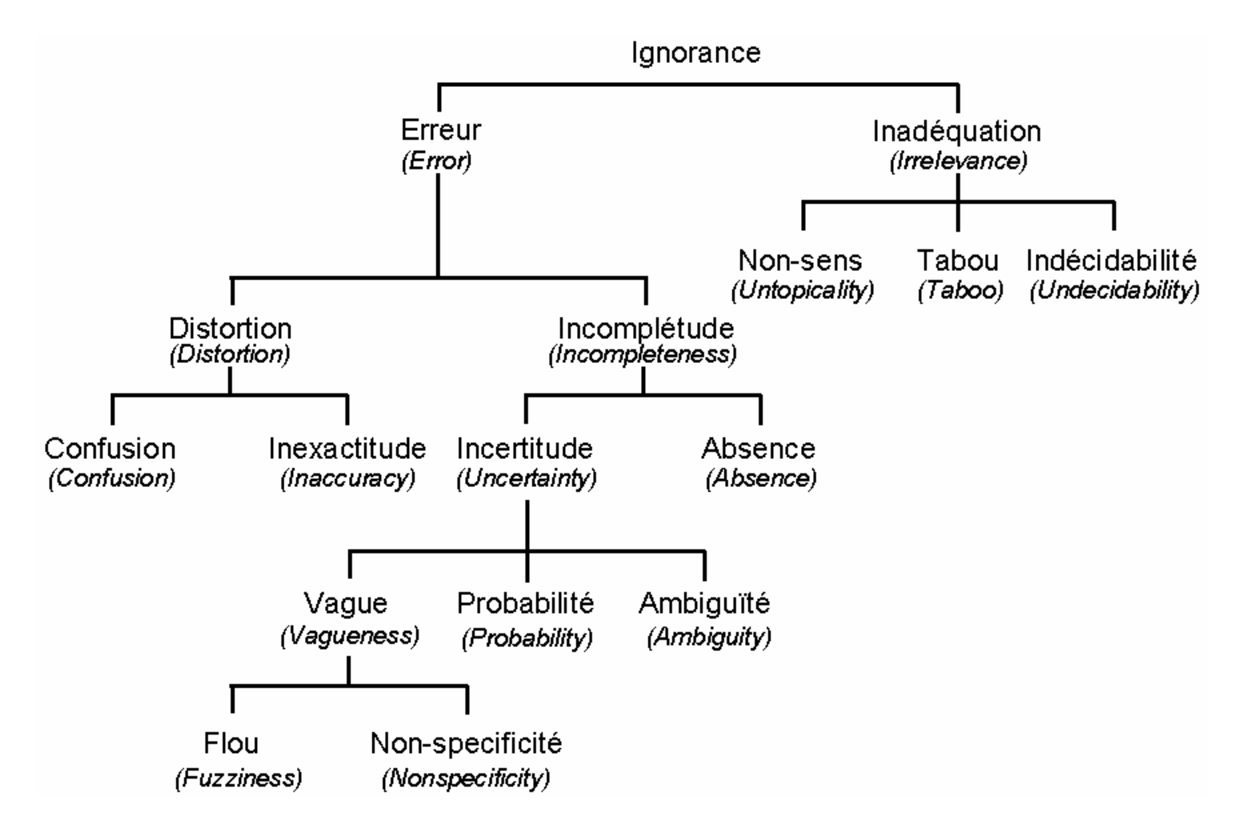


Figure 4: Taxonomie de l'ignorance (traduit de Smithson, 1989 - les termes originaux sont mis entre parenthèse en italique)

L'ignorance fait l'objet de réflexions depuis l'antiquité. Le premier grand philosophe Grec, Socrate (puis son disciple Platon), disait « je ne sais qu'une chose, c'est que je ne sais rien ». Le fait d'être conscient de son ignorance était pour lui en soi un signe de sagesse et un niveau plus élevé d'intelligence. Il distinguait le fait d'ignorer quelque chose du fait

d’ignorer mais en n’étant pas conscient de notre ignorance (voire de penser à tord que l’on sait), nommant le premier « simple ignorance » et le second « double ignorance ». Martinet et Marti (2001) les identifient comme « ignorance savante » et « ignorance profonde » et Smithson (1989) « ignorance consciente » et « méta-ignorance ». Bédard (1986), dans ses travaux sur les sources de distorsion de l'information, parle de méta-incertitude (c.à.d. l'incertitude sur l'incertitude) qu'il est important de bien connaître (c.à.d. sortir de la double ignorance) pour utiliser les données géospatiales en connaissance de cause.

Dans le domaine de l’information géographique, Fisher (1999) présente une taxonomie de l’incertitude (cf. Figure 5), formant un sous-ensemble de la taxonomie plus globale de Smithson. Il fait ensuite le lien entre les concepts (ex. vague) et les méthodes pouvant être utilisées pour gérer et représenter ces concepts (ex. théorie des ensembles flous), certains concepts n’ayant parfois pas de méthode associée. Dans cette classification, l’erreur est associée aux objets bien définis et peut être modélisée par des approches statistiques (probabilités). Les objets mal définis, fréquemment rencontrés dans le domaine des ressources naturelles (ex. limite d’une forêt, limite entre deux types de sols), peuvent eux être vagues (modélisé par des approches logiques telles que la théorie des ensemble flous) ou ambigus.

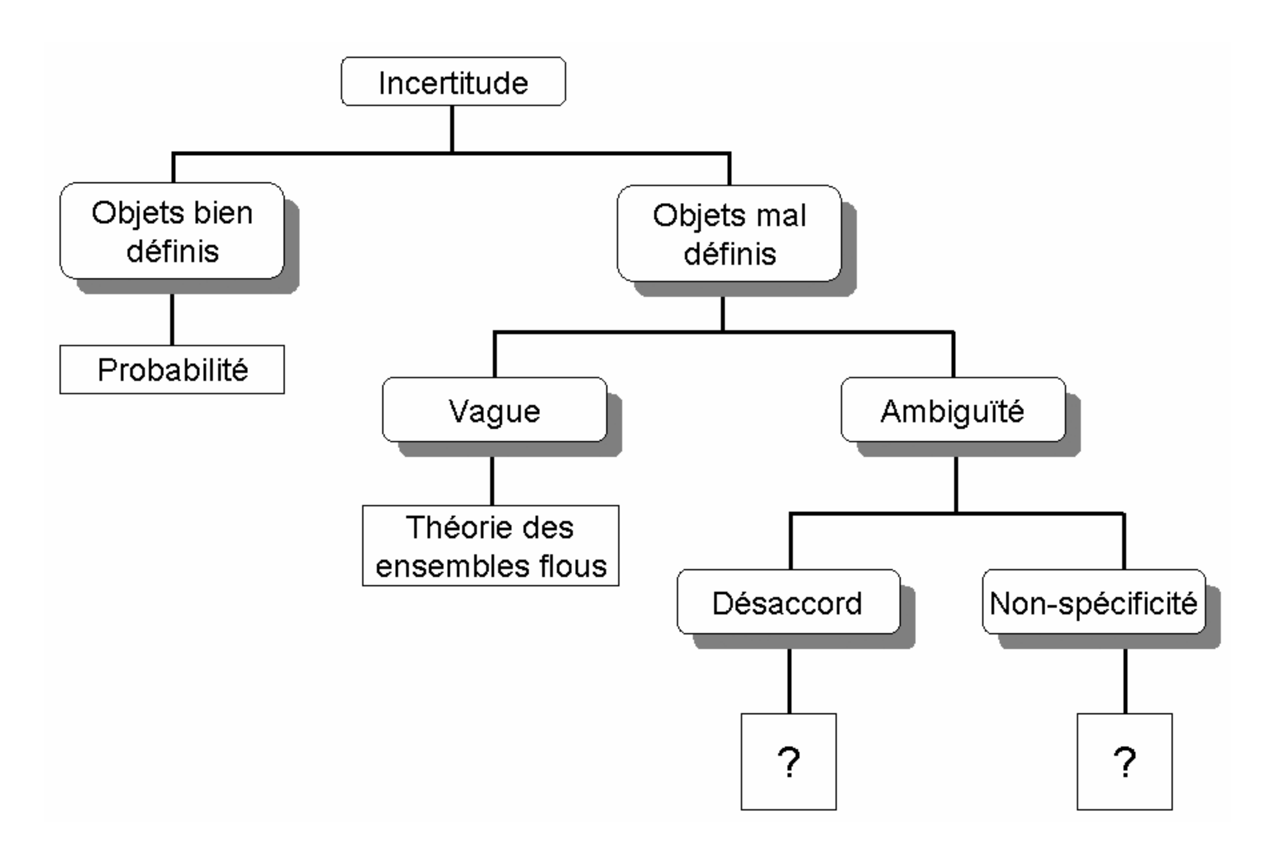


Figure 5: Taxonomie de l'incertitude (traduit de Fisher, 1999)

Goodchild *et al.* (1994) différencient l'incertitude de l'erreur, l'incertitude étant une mesure relative de la divergence, tandis que l'erreur est une valeur pour cette mesure. Windholz (2001) définit l'incertitude comme un état de connaissance sur les relations existant entre le monde et un état de ce monde.

Une différence doit être faite entre les termes précision et exactitude qui sont souvent utilisés indifféremment dans le langage courant. Le terme précision peut avoir deux définitions (Longley *et al.*, 2001). Une première réfère à la capacité qu'ont certains outils de mesure à obtenir des mesures successives les plus similaires possibles pour un même phénomène observé (ex. positions d'un même site enregistrées par GPS). La seconde, plus générale, se réfère au nombre de décimales fournies par un appareil de mesure; plus la mesure aura de décimales et plus elle sera précise. L'exactitude peut elle être définie comme la proximité d'une mesure par rapport à une valeur vraie ou considérée comme telle (David et Fasquel, 1997; Mowrer, 1999). Ainsi, des données géospatiales peuvent être enregistrées avec beaucoup de précision mais être totalement inexactes.

2.2.2 Concept de qualité

Le terme « qualité » vient du latin « *qualitas* », néologisme basé sur « *qualis* » signifiant « quel » (c.à.d. la nature d'une chose). On trouve encore ce sens dans l'expression « en qualité de », ainsi qu'en philosophie où la qualité peut être définie comme l'« aspect de l'expérience qui diffère spécifiquement de tout autre aspect et, par là, permet de distinguer cette expérience » (Office québécois de la langue française, 2004).

La qualité est une préoccupation que l'on retrouve dans beaucoup d'autres domaines que la géomatique. Dès l'antiquité, des philosophes grecs tels que Socrate, Platon et Aristote associaient la qualité à l'excellence. Dès le début du XXème siècle, différentes significations ont été associées au concept de qualité, issues principalement du domaine de la confection et de la distribution de produits et de services. Deux grands groupes de définitions peuvent ainsi être identifiés. Le premier associe la qualité d'un produit ou d'un service au respect de normes, spécifications, permettant d'élaborer des produits exempts d'erreurs (ex. Crosby, Lewitt, Gilmore). Le second associe la qualité à la satisfaction des utilisateurs utilisant ce produit ou service, un produit de qualité devant rencontrer ou excéder les besoins des utilisateurs (ex. Juran, Gronroos, Deming). Ces deux concepts sont fréquemment identifiés par « qualité interne » et « qualité externe » (Aalders, 2002; Dassonville *et al.*, 2002). En géomatique, la première vision se place généralement du point de vue des producteurs de données, comparativement à la seconde qui se place du point de vue des utilisateurs. Un produit est donc jugé de qualité pour les producteurs s'il est conforme à des spécifications définies, tandis qu'un produit est de qualité pour les utilisateurs s'il rencontre ou dépasse leurs attentes (Kahn et Strong, 1998). Juran et al. (1974) sont les premiers à définir la qualité par le concept d'adéquation à l'utilisation (*fitness for use*) largement utilisé en géomatique et adopté par les organismes internationaux comme définissant la qualité (ex. ISO, IEEE).

Le comité international de normalisation ISO (*International Standard Organization*) définit la qualité comme étant « l'adéquation aux exigences; satisfaction des besoins de l'utilisateur » et la qualité d'un produit comme « la totalité des caractéristiques d'un produit ou service qui influent sur sa capacité à satisfaire les besoins explicites ou implicites du client » (ISO 8402, 1994). La qualité étant l'adéquation à l'utilisation, un jeu de données ne

peut donc pas se voir attribuer une valeur unique de qualité, celle-ci pouvant varier d'un utilisateur à un autre ou également, pour un même utilisateur, d'une application à une autre. Un jeu de données ne peut donc pas avoir une qualité absolue étant donné qu'il est impossible de satisfaire les besoins de tous les types d'utilisateurs dans tous les contextes possibles.

Tandis que de nombreux travaux ont porté sur la définition de la qualité interne, encore peu d'études se sont penchées sur les problèmes de qualité externe. Parmi ces travaux, Wang et Strong (1996) classifient la qualité selon le point de vue des utilisateurs suivant plusieurs axes (dimensions). Se basant sur un sondage effectué auprès d'environ 350 utilisateurs de données, ils classifient la qualité suivant quatre dimensions:

- *Qualité intrinsèque* (crédibilité, précision, objectivité et réputation);
- *Qualité contextuelle* (valeur ajoutée, pertinence, à propos, complétude, volume de données approprié);
- *Qualité représentationnelle* (interprétabilité, facilité de compréhension, consistance de la représentation, concision de la représentation);
- *Accessibilité de la qualité* (accessibilité, sécurité d'accès).

Wang et Strong définissent le concept de dimension de la qualité comme « un ensemble d'attributs, définissant la qualité des données, qui représentent un aspect unique de la qualité des données » (traduction libre).

2.2.3 Qualité des données géospatiales

Les données géospatiales rencontrent en partie les mêmes problèmes que les données plus traditionnelles ou les produits, de manière plus générale, en regard de la qualité. Les problèmes de documentation de la qualité ont connu un intérêt croissant lors de la dernière décennie, entre autres en raison de l'accroissement de la diffusion des données entre organisations (Goodchild, 1995; Chrisman, 1999; Veregin, 1999).

Le concept d'adéquation à l'utilisation (*fitness for use*), introduit en 1982 par la norme américaine NCDCDS et par Chrisman (1983) dans la communauté de l'information géographique est aussi maintenant largement adopté par cette communauté comme

définissant le concept de qualité (Veregin, 1999). Cependant, l'utilisation du concept de qualité dans les travaux scientifiques est souvent contradictoire, le concept étant souvent défini dans un premier temps par *fitness for use*, puis employé par la suite en ne considérant que le seul aspect de précision spatiale. Le comité de normalisation en géomatique ISO/TC 211 reprend pour les données géographiques la même définition générale de la qualité donnée par l'ISO 9000. Bédard et Vallière (1995) précisent cette définition en y ajoutant le contexte d'utilisation, définissant la qualité comme étant « l'ensemble des caractéristiques qui la rendent [la donnée à référence spatiale] apte à satisfaire les besoins définis par un utilisateur dans le cadre d'une application précise ».

La dualité de point de vue entre producteurs et utilisateurs de données vis à vis du concept de qualité apparaît également dans le domaine spatial (Frank, 1998; Tastan et Altan, 1999). Bien que la qualité soit définie par le concept de *fitness for use*, les producteurs utilisent en général le concept de qualité pour la seule qualité interne et nomment *fitness for use* la qualité externe (*cf.* Figure 6).

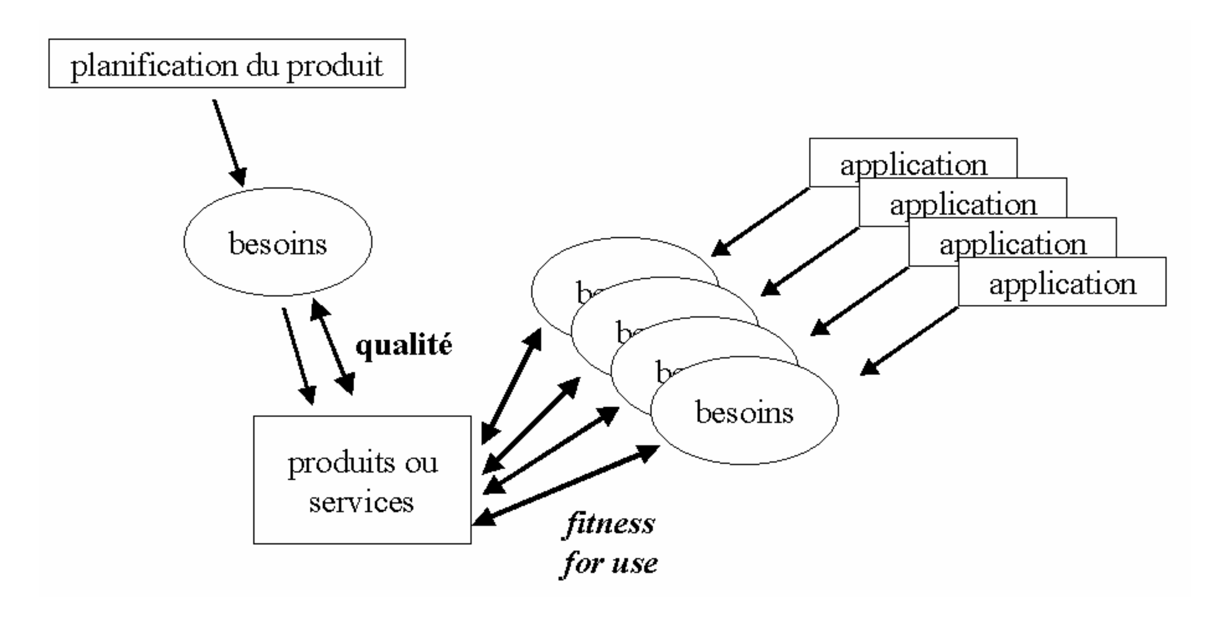


Figure 6: Concepts de qualité interne et externe (*fitness for use*) des données (traduit de Morrisson, 1995)

La Figure 6 illustre clairement l'opposition des deux points de vue, la qualité pour le producteur étant vue d'un côté comme le niveau de similarité entre la représentation de la

réalité désirée (terrain nominal) et le jeu de données effectivement produit, et de l'autre comme l'adéquation entre les produits et services et les besoins des utilisateurs en fonction d'une ou plusieurs applications données.

Dans le domaine académique, beaucoup de travaux de recherche actuels traitant du domaine de la qualité des données géospatiales s'intéressent à la caractérisation de l'incertitude spatiale et la modélisation de sa propagation (Lowell et Jaton, 1999; Heuvelink et Lemmens, 2000; Hunter et Lowell, 2002). Ces approches, en général basés sur des approches quantitatives (ex. simulation Monte-Carlo), ne sont souvent utilisables que dans des cas précis (Morrison, 1995; Lowell et Jaton, 1999). Ces travaux ont typiquement des approches de type « producteur de données » quant à la perception de la qualité. Il semble toutefois y avoir un intérêt croissant pour intégrer ce type d'approche au sein des processus de prise de décision (Lowell, 2004).

Bédard et Vallière (1995) soutiennent qu'il « n'existe pas et n'existera jamais de méthode générique rigoureuse, mathématique, permettant de calculer de façon parfaitement objective la qualité de n'importe quelle donnée à référence spatiale. Il demeure tout de même possible de mesurer la qualité avec des indicateurs tant qualitatifs que quantitatifs, et ceci, de manière utile » et qu'il est « possible d'utiliser un ensemble minimal de critères » permettant de décrire la qualité. Plusieurs auteurs, tel que Kahn et Strong (1998), soutiennent que la qualité du point de vue de l'utilisateur doit rencontrer ou dépasser les besoins de l'utilisateur. Bédard et Vallière proposent dans ce sens un système d'évaluation de la qualité d'un jeu de données pour lequel le jeu de données de qualité rencontre les besoins des utilisateurs (sans les dépasser). Le *Center for Technology in Government* (CTG, 2000) souligne la nuance à faire entre des données parfaites et des données adéquates pour l'utilisation, rejoignant le concept de *satisfycing* de Simon (1955) bien connu dans le domaine de la prise de décision. Des considérations économiques entrent alors en ligne de compte, l'utilisateur devant faire un compromis entre le coût des données et leur qualité (Charron, 1995; Holmwood, 2000).

Plusieurs auteurs décomposent le concept de qualité en sous-classes. Veregin (1999) définit trois composantes pour la qualité des données géospatiales: la position, le temps et le thème, classification inspirée des travaux de Berry (1964) et Sinton (1978). Il associe ces

axes à la précision et la résolution (précision spatiale, temporelle et thématique, etc.). Bédard et Vallière (1995) proposent six caractéristiques permettant de définir la qualité d'un jeu de données spatial:

- *Définition* : Permet d'évaluer si la nature exacte d'une donnée et de l'objet qu'elle décrit, c.à.d. le « quoi », correspond aux besoins (définitions sémantique, spatiale et temporelle);
- *Couverture* : Permet d'évaluer si le territoire et la période pour lesquels la donnée existe, c.à.d. le « où » et le « quand », correspondent aux besoins;
- *Généalogie* : Permet de connaître d'où provient une donnée, ses objectifs d'acquisition, les méthodes utilisées pour l'obtenir, c.à.d. le « comment » et le « pourquoi », et de voir si cela correspond aux besoins;
- *Précision* : Permet d'évaluer ce que vaut une donnée et si elle est acceptable pour le besoin exprimé (précision sémantique, temporelle et spatiale de l'objet et ses attributs);
- *Légitimité* : Permet d'évaluer la reconnaissance officielle et la portée légale d'une donnée et si elles rencontrent les besoins (standards de facto, respect de normes reconnues, reconnaissance légale ou administrative par un organisme officiel, garantie légale par un fournisseur, etc.);
- *Accessibilité* : Permet d'évaluer la facilité avec laquelle l'usager peut obtenir la donnée analysée (coût, délai, format, confidentialité, respect des normes reconnues, droits d'auteur, etc.).

2.3 Documentation et communication de la qualité

Différentes étapes doivent être effectuées avant de pouvoir utiliser des informations sur la qualité des données au sein d'un processus de prise de décision. Cette section présente succinctement ces étapes, soit (1) l'évaluation de la qualité des données, (2) la gestion des informations décrivant la qualité des données puis (3) la communication de ces informations aux utilisateurs des données.

2.3.1 Évaluation et documentation de la qualité interne

Afin d'évaluer la qualité interne de jeux de données, les producteurs de données doivent comparer les données produites aux données qui auraient dû être produites (c.à.d. données produites sans erreurs) (cf. Figure 7). Ces données idéales sont souvent nommées « terrain nominal » ou « univers du discours », le terrain nominal étant défini par David et Fasquel (1997) comme une « image de l'univers, à une date donnée, à travers le filtre défini par les spécifications de produit ». Toutefois, comme le terrain nominal n'est pas un jeu de données avec une existence physique réelle, il est remplacé par un jeu de données de référence (aussi nommé « données de contrôle »), plus exact que le jeu de données produit (David et Fasquel, 1997).

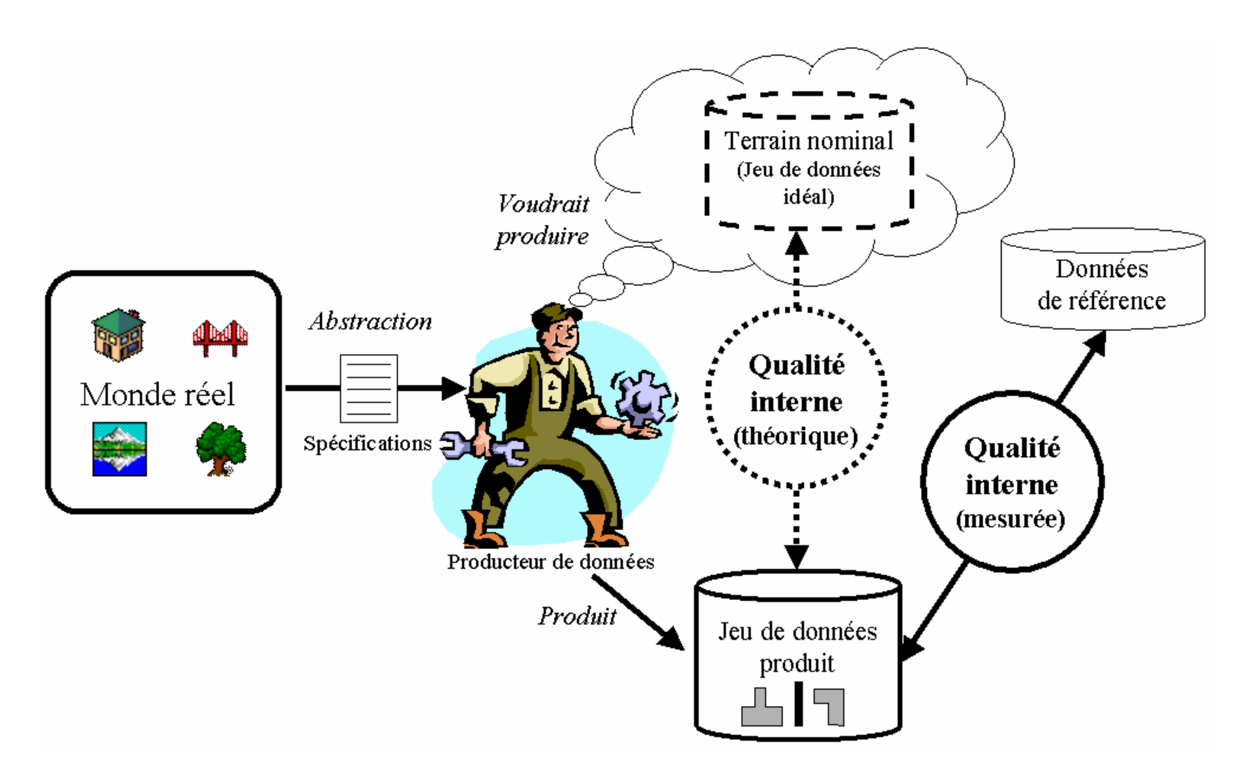


Figure 7 : Concepts de qualité interne et son évaluation

L'évaluation de la qualité interne consiste alors dans l'identification des objets représentant les mêmes phénomènes dans les deux jeux de données (c.à.d. processus d'appariement) pour ensuite les comparer pour un ensemble de critères reliés par exemple aux composantes

spatiales, sémantiques et temporelles. La Figure 8 présente la dualité entre producteurs et utilisateurs de données. L'univers du discours (*Universe of discourse*) étant défini par l'ISO comme « une vue du monde réel ou hypothétique incluant tous les éléments d'intérêt » (traduction libre) (ISO-TC/211, 2002).

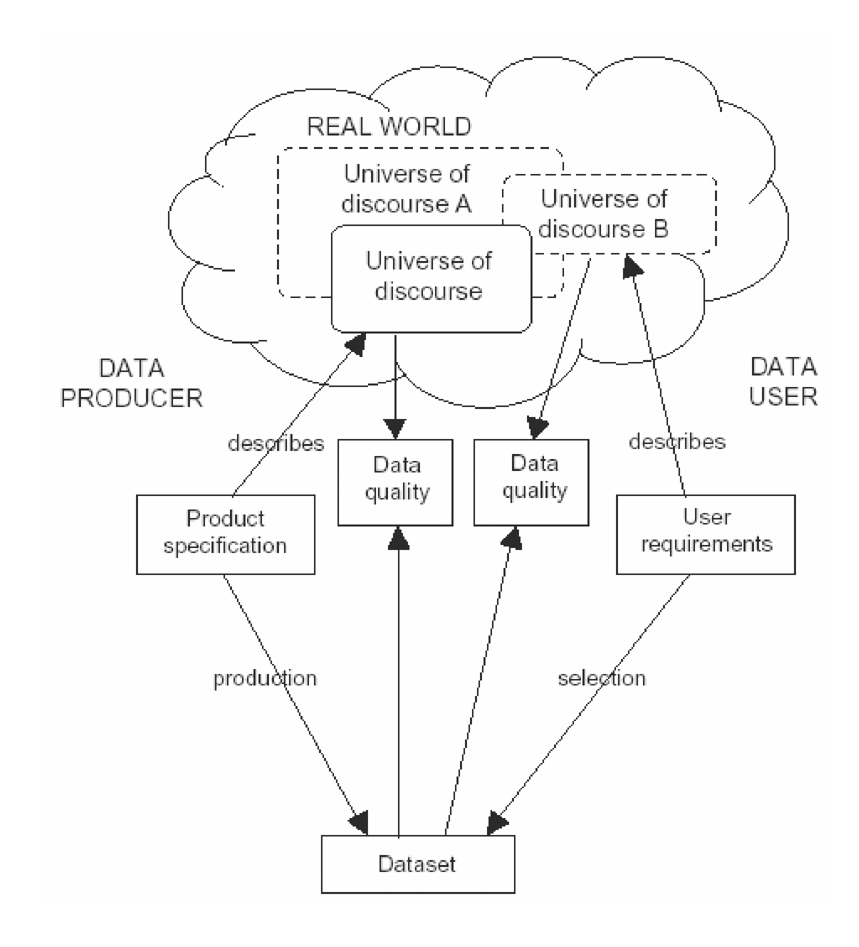


Figure 8: Cadre conceptuel pour la définition de la qualité (ISO-TC/211, 2002)

Étant donné le grand volume d'information à comparer, le temps de traitement qui serait nécessaire pour évaluer la qualité de chaque donnée, la disponibilité d'autres données suffisamment exactes pour permettre une comparaison, etc., les producteurs de données procèdent généralement à un échantillonnage, jugé représentatif de l'ensemble des données, allant permettre d'évaluer la qualité de l'ensemble des données produites. Différents types d'échantillonnage peuvent être utilisés, dépendant entre autres de la taille et de la distribution de la population globale (Faïz, 1999; ISO-TC/211, 2002).

Il est difficile, voire impossible, de caractériser la qualité d'un jeu de données à l'aide d'un critère unique. Ainsi, différents aspects peuvent être analysés. Bien qu'il existe une certaine variabilité dans les critères de qualité utilisés pour décrire des données géospatiales, on retrouve depuis une vingtaine d'années un ensemble de critères présents dans la plupart des procédures d'estimation de la qualité et des normes décrivant la qualité (ex. CEN, ISO, FGDC, IGN). Ces critères sont: la généalogie des données, l'exactitude spatiale, l'exactitude sémantique, l'exactitude temporelle, la complétude et la cohérence logique (cf. chapitre 4 pour plus de détails). Chacun de ces critères est par la suite composé de sous-critères.

Pour chacun de ces critères, des mesures sont faites, pour les échantillons sélectionnés, entre les données produites et les données de contrôle. Il existe une grande variété de techniques permettant de mesurer les différents critères de qualité (ex. moyenne quadratique des erreurs pour la précision géométrique ou taux de confusion pour la précision sémantique) (David et Fasquel, 1997).

2.3.2 Gestion de l'information sur la qualité

Les informations sur la qualité peuvent décrire la qualité de données à différents niveaux de détails. Certaines informations peuvent par exemple être associées à un jeu de données dans sa globalité (c.à.d. ensemble des objets le composant), d'autres peuvent par exemple porter sur une classe d'objets spécifique (ex. uniquement les routes) ou encore sur une instance d'objet particulière. Hunter (2001) identifie la granularité des informations sur la qualité comme devant être une des considérations principales des travaux de recherche futurs portant sur la qualité des données géospatiales. Il mentionne que « la qualité souffre en général d'une représentation faite à un niveau trop général plutôt qu'à des niveaux de granularité plus fins » (traduction libre). Hunter fournit plusieurs exemples de métadonnées actuelles montrant les limites d'une représentation trop générale des métadonnées, telles que: l'exactitude spatiale est « variable », « de 100m à 1000m » ou encore « +/- 1.5m (urbain) à +/- 250m (rural) ». Ces exemples illustrent le fait que l'hétérogénéité de la qualité des données géospatiales n'est pas suffisamment documentée dans les métadonnées actuelles, ne permettant pas, par exemple, de connaître la qualité d'un sous-ensemble du jeu de données, d'un objet en particulier, etc. De plus, Hunter mentionne que la documentation

de la qualité à un niveau trop agrégé ne permet pas d'avoir une connaissance de la variation spatiale de la qualité, bien que cette information serait utile aux utilisateurs.

Plusieurs auteurs se sont intéressés à la manière de gérer cette granularité de métadonnées, proposant différents modèles.

Faïz (1996 et 1999) présente une méthode permettant de gérer et de communiquer l'information sur la qualité à différents niveaux de détails, basés sur une structure de données relationnelle avec les SIG GEO_2 et ArcInfo. Il utilise cinq niveaux de détails: base de données, couche de données, objet complexe, objet simple et les coordonnées. Son approche a principalement pour objectif de fournir des informations sur la qualité aux producteurs de données (ex. IGN France) pour leur permettre d'identifier les erreurs de leurs produits et ainsi améliorer la qualité interne des données produites.

Qiu et Hunter (1999 et 2002) présentent eux aussi un modèle permettant la gestion de métadonnées sur la qualité à différents niveaux de détails. Se basant sur la base de données topographique australienne au 250K, ils identifient quatre niveaux de détails: *data set*, *data layer*, *feature class* et *feature*. Dans ce modèle, chaque objet de niveau détaillé hérite des attributs de ses parents (héritage en Orienté-Objet). Les auteurs présentent un prototype implémentant leurs concepts en couplant la base de données MS-Access et le SIG ArcView, permettant ainsi le stockage, l'accès, la mise à jour, et la visualisation des informations sur la qualité.

Bédard et Vallière (1995) proposent une méthode permettant d'agréger six caractéristiques décrivant la qualité de données (attribut, géométrie et existence) en instances d'objets, classes et jeux de données.

La norme 19114 de l'ISO/TC 211 (2003) propose un cadre général pour encoder les métadonnées dans un but de recherche, d'échange et de présentation des métadonnées. Ils proposent une hiérarchie pouvant être utilisée pour stocker les métadonnées à différents niveaux de détails. Cette hiérarchie peut aider à filtrer ou préciser des requêtes des utilisateurs pour un niveau de détail désiré. La hiérarchie ISO va plus loin que celles de Faïz ou Qiu et Hunter en permettant d'associer des métadonnées aux attributs (attributs et instances). Les niveaux de métadonnées de l'ISO 19114 sont: *data series*, *dataset*, *feature type*, *feature instance*, *attribute type* et *attribute instance*.

Des hiérarchies peuvent également être retrouvées dans les organismes produisant des données géospatiales. Par exemple, les métadonnées de la Base Nationale de Données Topographique du Canada (BNDT) sont communiquées dans un fichier texte fourni avec le jeu de données numérique. Les métadonnées de ce fichier sont réparties en cinq sections:

- *Territoire* (ex. numéro du feuillet, nom du jeu de données, province, zone de projection);
- *Jeu de données* (ex. date à laquelle le jeu de données a été rendu disponible dans la BNDT);
- *Intégration* (ex. pourcentage d'intégration validé entre des feuillets cartographiques adjacents);
- *Polygone*: Métadonnées communes à l'ensemble des objets situés dans une certaine zone définie par des coordonnées géographiques (ex. type de méthode d'acquisition). Chaque jeu de données peut inclure un à plusieurs polygones de métadonnées;
- *Thème*: Métadonnées reliées à un thème en particulier (ex. nom du thème, disponibilité, résolution).

Certaines métadonnées de la BNDT sont également reliées aux primitives géométriques (ex. exactitude spatiale). Ces métadonnées ne sont pas inclues dans le fichier texte mais directement stockées comme des attributs dans le fichier de données. Ainsi, les métadonnées de la BNDT possèdent quatre niveaux de détails: jeu de données (les sections territoire et intégration sont aussi associées au niveau du jeu de données), polygone de métadonnées, thème et primitives géométriques.

Les informations décrivant la qualité, incluses dans les métadonnées de la base de données topographique Australienne (250K), sont aussi documentées à quatre niveaux de détails: *dataset*, *data layer*, *feature class* et *individual feature level* (Hunter, 2001).

2.3.3 Communication et utilisation de l'information sur la qualité

L'information sur la qualité a pour objectif de permettre aux utilisateurs de déterminer dans quelle mesure les données répondent à leurs besoins (concept de *fitness for use*) (Chrisman,

1990; Agumya et Hunter, 1997). Pour cela, différentes manières de communiquer l'information sur la qualité sont utilisées ou proposées dans la littérature.

Le moyen le plus utilisé actuellement pour communiquer l'information sur la qualité est la diffusion de métadonnées, incluant certaines informations sur la qualité. Les organismes de normalisation suggèrent l'inclusion d'informations décrivant la qualité des jeux de données (ex. ISO 19113 et ISO 19115; FGDC[15], CEN[16]). Toutefois, l'utilité de ces métadonnées reste très limitée étant donné, entre autres, la complexité de leur représentation, celles-ci étant même difficiles à comprendre pour des experts en géomatique (Gervais, 2004). Hunter et Masters (2000) mentionnent même que les informations fournies par les producteurs sur la qualité sont de plus en plus perçues par les utilisateurs comme étant uniquement un moyen pour les producteurs de se couvrir en cas de litiges possibles.

Étant donné les limitations des métadonnées dans leur format actuel, certaines recherches ont exploré des façons de visualiser l'information sur la qualité. De nombreux travaux portant sur la visualisation de l'information sur la qualité ont été effectués, notamment dans le cadre de l'initiative de recherche nº 7 du NCGIA, « *Visualizing the Quality of Spatial Information* », dirigée par K. Beard et B. Buttenfield entre 1991 et 1993 (Buttenfield et Beard, 1991; Beard et Mackaness, 1993; Buttenfield, 1993; McGranaghan, 1993; Buttenfield et Beard, 1994; Fisher, 1994a; Goodchild *et al.*, 1994; Faïz, 1996; Beard, 1997; Beard et Buttenfield, 1999; Leitner et Buttenfield, 2000; Windholz, 2001; Drecki, 2002). Ces travaux proposent diverses méthodes permettant de représenter les différents critères de qualité (ex. exactitude spatiale, complétude, cohérence logique) pour différentes primitives géométriques (ex. points, lignes, polygones). Dans le domaine de la représentation graphique, l'ouvrage de référence de Bertin (1973) sur la sémiologie graphique, identifie six variables visuelles: la taille, la valeur, la couleur, l'orientation, la forme et la texture. Ces variables ont ensuite été étendues par différents auteurs, ajoutant par exemple la saturation des couleurs et la clarté/focus (Morrison, 1974; Mac Eachren, 1992; McGranaghan, 1993). Chacune de ces variables peut être utilisée lorsque l'on représente des informations géospatiales et beaucoup de méthodes visualisant la qualité se basent sur ces variables (ex. changements de couleur ou de texture des objets en fonction de leur

[15] Federal Geographic Data Committee (http://www.fgdc.gov)
[16] Comité Européen de Normalisation (http://www.cenorm.be)

qualité). On retrouve également beaucoup d'autres méthodes telles que la représentation floue des objets, la visualisation de surfaces 3D représentant la variabilité spatiale de la qualité, l'implantation de filtres ne sélectionnant que les objets ayant un certain niveau de qualité, etc. (Paradis et Beard, 1994; Beard et Buttenfield, 1999). Certains travaux exploitent également la diffusion d'information sonores ou d'animations (Fisher, 1994b).

D'autres travaux complémentaires visent à exploiter les informations sur la qualité des données dans les logiciels actuels (ex. SIG) afin, entre autres, de limiter les risques de mauvaise utilisation de la part des usagers. Ces méthodes nécessitent des structures de données permettant de gérer les informations sur la qualité (cf. section précédente). Reinke et Hunter (2002) présentent un modèle de communication de l'incertitude, adapté de Gottsegen *et al.* (1999) (cf. Figure 9). Dans ce modèle, la représentation est centrale dans le processus de communication de l'incertitude. Toutefois, cette représentation ne se fait pas de manière unidirectionnelle (c.à.d. du système à l'utilisateur), mais suggère des rétroactions entre les deux, permettant ainsi une plus grande interaction de l'usager pouvant mener à une meilleure communication.

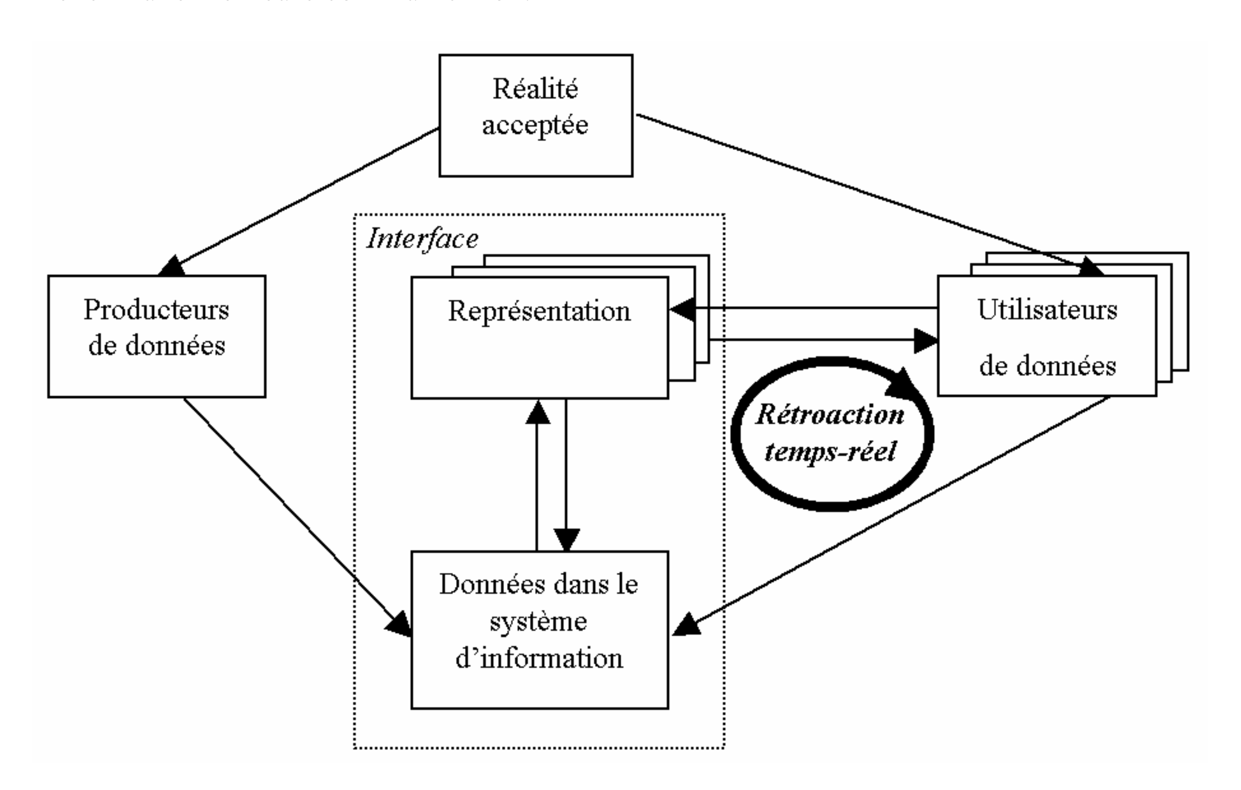


Figure 9: Modèle de communication aux usagers de l'incertitude dans les bases de données géospatiales (traduit de Reinke et Hunter, 2002)

Basé sur ce modèle, Reinke et Hunter proposent des bases théoriques permettant de communiquer l'incertitude des données aux utilisateurs. Cette communication, utilisant des informations sur la qualité stockées dans une base de données, peut par exemple être faite sous la forme de messages faits aux utilisateurs émis lorsqu'ils effectuent des opérations jugées illogiques en fonction des données utilisées et de leur qualité (Beard, 1989; Hunter et Reinke, 2000).

D'autres travaux visent de manière plus générale à développer des SIG prenant en compte les incertitudes dans différentes opérations (ex. précision des résultats des opérateurs du SIG, propagation d'erreur lors de mises à jour). Il y a plus de 10 ans déjà, Burrough (1992) parlait de développer des « SIG intelligents » allant mettre à profit les métadonnées disponibles pour guider les utilisateurs dans l'utilisation de données entachées d'incertitude. Unwin (1995) amène le concept de « *error-sentive GIS* » qui désigne un SIG offrant des fonctionnalités de base pour la gestion des erreurs. Duckham et McCreadie (1999 et 2002) proposent le terme « *error-aware GIS* » offrant des extensions à l'*error-sentive GIS* pour des applications particulières et en utilisant des méthodes avancées provenant de l'intelligence artificielle ou des bases de données.

2.4 Outils d'intelligence décisionnelle

Certains outils provenant du domaine de l'intelligence décisionnelle ont été explorés puis exploités afin de permettre la gestion et la communication des informations sur la qualité visée par cette thèse.

Pour les aspects de communication des informations sur la qualité, les tableaux de bord de gestion et l'utilisation d'indicateurs ont été explorés. Cette exploration et sa revue de littérature sont présentées dans la section 3.4.

Pour les aspects concernant la gestion des informations sur la qualité, une exploration des bases de données multidimensionnelles et des outils SOLAP a été faite. Ces outils permettent typiquement de gérer des informations à différents niveaux de détails comme requis par les informations décrivant la qualité des données. Cette exploration et sa revue

de littérature sont présentées dans la section 4.7, au sein de l'article portant sur la gestion des informations sur la qualité.

2.5 Synthèse

En résumé, basé sur ces revues de littérature, nous avons constaté dans la section 2.1 que l'information géographique n'est jamais conforme à la réalité car (1) elle n'est qu'un modèle de cette réalité et (2) elle est toujours entachée d'erreurs (formant l'incertitude de manière générale). Les utilisateurs doivent alors pouvoir comprendre l'incertitude reliée aux données pour intégrer cette connaissance dans leur processus de prise de décision plus global. Nous avons vu que ces informations doivent être communiquées aux usagers dans un langage le plus proche possible du leur pour que le processus de communication soit efficace. Dans la section 2.2, nous avons introduit les concepts de qualité et les autres concepts connexes. Nous avons distingué les concepts de qualité interne et de qualité externe. La connaissance de l'incertitude permet aux usagers d'évaluer la qualité externe des données utilisées. L'évaluation de cette qualité permet de réduire les risques de mauvaise utilisation des données et ainsi réduire les risques de conséquences néfastes pouvant découler de cette mauvaise utilisation. Dans la section 2.3, nous avons décrit différentes étapes que suivent les informations sur la qualité, de l'évaluation de la qualité interne menant à la production de métadonnées, à leur gestion puis leur communication. Nous avons ainsi montré que les informations fournies par les producteurs de données (c.à.d. métadonnées) sont, dans leur forme actuelle, d'une aide très limitée, mais qu'elles peuvent servir de base à des méthodes plus efficaces de communication des informations sur la qualité. Nous avons montré que pour cela, ces informations devraient être stockées à différents niveaux de détails afin de préserver leur richesse et qu'elles doivent par la suite être communiquées aux usagers sous la forme de représentations intuitives et permettant aux usagers d'interagir avec le système. Finalement, nous avons présenté dans la section 2.4 un aperçu de méthodes allant être utilisées dans cette thèse, soit les tableaux de bord et les indicateurs pour la communication de l'information sur la qualité, et les bases de données multidimensionnelles et les outils SOLAP pour la gestion de ces informations.

2.6 Références

Aalders H.J.G.L., "The Registration of Quality in a GIS". *Spatial Data Quality* (W. Shi, P. Fisher, et M.F. Goodchild, Eds), Taylor & Francis, p. 186-199, 2002.

Agumya A., Hunter G.J., "Determining fitness for use of geographic information", *ITC Journal*, vol. 2, nº 1, 1997, p. 109-113.

Beard K., "Use error: the neglected error component", *Proceedings of AUTO-CARTO 9*, Baltimore, Maryland, March 1989, p. 808-817.

Beard K., "Representations of Data Quality". *Geographic Information Research: Bridging the Atlantic* (M. Craglia et H. Couclelis, Eds), Taylor and Francis, p. 280-294, 1997.

Beard K., Buttenfield B., "Detecting and evaluating errors by graphical methods". *Geographical Information Systems* (P.A. Longley, M.F. Goodchild, D.J. Maguire et D.W. Rhind, Eds), Wiley, p. 219-233, 1999.

Beard K., Mackaness W., "Visual Access to Data Quality in Geographic Information Systems", *Cartographica*, vol. 30, nº 2-3, 1993, p. 37-45.

Bédard Y., *A Study of the Nature of Data Using a Communication-Based Conceptual Framework of Land Information Systems*, Thèse de doctorat, University of Maine, Orono (USA), 1986.

Bédard Y., "Uncertainties in Land Information Systems Databases", *Proceedings of Eighth International Symposium on Computer-Assisted Cartography*, Baltimore, Maryland (USA), 29 Mars - 3 Avril 1987, American Society for Photogrammetry and Remote Sensing and American Congress on Surveying and Mapping, p. 175-184.

Bédard Y., Vallière D., 1995. *Qualité des données à référence spatiale dans un contexte gouvernemental*. Rapport de recherche sur la mise en place d'une méthode d'évaluation de la qualité des données à référence spatiale préparé pour le Plan géomatique du Gouvernement du Québec, Université Laval, Québec, Canada.

Berry B., "Approaches to regional analysis: a synthesis." *Annals of the Association of American Geographers*, vol. 54, 1964, p. 2-11.

Bertin J., *Sémiologie graphique: les diagrammes, les résaux, les cartes*, Paris, Mouton-Gauthier-Villars-Bordas, 1973.

Box G.E.P., "Science and statistics", *Journal of the American Statistical Association*, vol. 71, 1976, p. 791-799.

Brodeur J. Bédard Y. Edwards G., Moulin B., "Revisiting the Concept of Geospatial Data Interoperability within the Scope of Human Communication Processes", *Transactions in GIS*, vol. 7, nº 2, 2003, p. 243-265.

Burrough P. A., "Development of intelligent geographical information systems", *International Journal of Geographical Information Systems*, vol. 6, nº 1, 1992, p. 1-11.

Buttenfield B., Beard K.M., "Graphical and Geographical components of Data Quality". *Visualization in Geographic Information Systems* (H. M. Hearnshaw, et D. J. Unwin, Eds), Wiley, p. 150-157, 1994.

Buttenfield B.P., "Representing Data Quality", *Cartographica*, vol. 30, nº 2-3, 1993, p. 1-7.

Buttenfield B.P., Beard K., "Visualizing the quality of spatial information", *Proceedings of AUTO-CARTO 10*, 1991, p. 423-427.

Charron J., *Développement d'un processus de sélection des meilleures Sources de données cartographiques pour leur intégration à une base de données à référence spatiale*, Mémoire, Université Laval, Québec, 1995.

Chrisman N.R., "The Role of Quality information in the Long Term Functioning of a Geographical Information System." *Proceedings of International Symposium on Automated Cartography (Auto Carto 6)*, Ottawa, Canada, 1983, p. 303-321.

Chrisman N.R., "The error component in spatial data". *Geographic Information Systems: Principles and Applications* (D. J. Maguire, M. F. Goodchild, et D. W. Rhind, Eds), Wiley, London, p. 165-174, 1990.

Chrisman N.R., "Speaking Truth to Power: An Agenda for Change". *Spatial Accuracy Assessment, Land Information Uncertainty in Natural Ressources* (K. Lowell, et A. Jaton, Eds), Quebec, p. 27-31, 1999.

CTG, 2000. Insider's Guide to Using Information in Government - The devil is in the data, Center for Technology in Government, http://www3.ctg.albany.edu/static/usinginfo/Data/data.htm.

Dassonville L. Vauglin F. Jakobsson A., Luzet C., "Quality Management, Data Quality and Users, Metadata for Geographical Information". *Spatial Data Quality* (W. Shi, P. Fisher, and M. F. Goodchild, Eds), Taylor & Francis, p. 202-215, 2002.

David B., Fasquel P., 1997. Bulletin d'information de l'IGN - Qualité d'une base de données géographique: concepts et terminologie, N. 67, IGN France.

De Bruin S. Bregt A., Van de Ven M., "Assessing fitness for use: the expected value of spatial data sets", *International Journal of Geographical Information Science*, vol. 15, nº 5, 2001, p. 457-471.

Drecki I., "Visualisation of Uncertainty in Geographic Data". *Spatial Data Quality* (W. Shi, P.F. Fisher and M.F. Goodchild, Eds), Taylor & Francis, p. 140-159, 2002.

Duckham M., McCreadie J., "An intelligent, distributed, error-aware OOGIS", *Proceedings of 1st International Symposium on Spatial Data Quality*, Hong Kong, 18-20 July 1999, p. 496-506.

Duckham M., McCreadie J. E., "Error-aware GIS Development". *Spatial Data Quality* (W. Shi, P. F. Fisher, and M. F. Goodchild, Eds), Taylor & Francis, London, p. 63-75, 2002.

Eco U., "De l'impossibilité d'établir une carte de l'empire à l'échelle de 1/1". *Pastiches et Postiches* (U. Eco, Eds), Éditions 10/18, p. 183, 2000.

Epstein E. F. Hunter G. J., Agumya A., "Liability insurance and the use of geographical information", *International Journal of Geographical Information Science*, vol. 12, nº 3, 1998, p. 203-214.

Faïz S. O., *Modélisation, exploitation et visualisation de l'information qualité dans les bases de données géographique*, Thèse de doctorat, Université Paris-Sud, Paris, 1996.

Faïz S. O., *Systèmes d'Informations Géographiques: Information Qualité et Data Mining*, Tunis, Editions C.L.E, 1999.

Fisher P., "Animation and sound for the visualization of uncertain spatial information". *Visualization in Geographic Information Systems* (H. M. Hearnshaw, and D. J. Unwin, Eds), Wiley, p. 181-185, 1994a.

Fisher P., "Visualising the uncertainty of soil maps by animation", *Cartographica*, vol. 30, 1994b, p. 20-27.

Fisher P. F., "Models of uncertainty in spatial data". *Geographical Information Systems* (P. A. Longley, M. F. Goodchild, D. J. Maguire, and D. W. Rhind, Eds), John Wiley & Sons, New-York, p. 191-205, 1999.

Frank A. U., "Metamodels for Data Quality Description". *Data Quality in Geographic Information - From Error to Uncertainty* (M. F. Goodchild, and R. Jeansoulin, Eds), Editions Hermes, p. 192, 1998.

Gervais M., *Pertinence d'un manuel d'instructions au sein d'une stratégie de gestion du risque juridique découlant de la fourniture de données géographiques numériques*, Thèse de doctorat, Département des Sciences Géomatiques, Université Laval, Québec, 2004.

Goodchild M. F., "Sharing Imperfect Data". *Sharing Geographic Information* (H. J. Onsrud, and G. Rushton, Eds), Rutgers University Press, New Brunswick, NJ, p. 413-425, 1995.

Goodchild M. F. Buttenfield B., Wood J., "Introduction to visualizing data validity". *Visualization in Geographic Information Systems* (H. M. Hearnshaw, and D. J. Unwin, Eds), Wiley, p. 141-149, 1994.

Gottsegen J. Montello D., Goodchild M. F., "A Comprehensive Model of Uncertainty in Spatial Data", *Proceedings of Spatial Accuracy Assessment: Land Information Uncertainty in Natural Resources*, Québec, Canada, Ann Arbor Press, 1998, p. 175-182.

Heuvelink G. B. M., Lemmens M. J. P. M., *4th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, Amsterdam, The Nederland, 2000.

Holmwood T. S., "Data Quality: Defining an achievable standard", *Proceedings of GITA Annual conference*, Denver (Colorado), USA, 2000.

Hunter G. J., "Managing uncertainty in GIS". *Geographical Information Systems* (P. A. Longley, M. F. Goodchild, D. J. Maguire, and D. W. Rhind, Eds), John Wiley & Sons, Inc., p. 633-641, 1999.

Hunter G. J., "Spatial Data Quality Revisited", *Proceedings of GeoInfo 2001*, Rio de Janeiro, Brazil, 4-5th October 2001, p. 1-7.

Hunter G. J., Lowell K., *5th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, Melbourne, Australia, 2002.

Hunter G. J., Masters E., "What's Wrong with Data Quality Information?" *Proceedings of GIScience 2000*, Savannah, USA, p. 201-203.

Hunter G. J., Reinke K. J., "Adapting Spatial Databases to Reduce Information Misuse Through Illogical Operations", *Proceedings of 4th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences (Accuracy 2000)*, Amsterdam, July 2000, p. 313-319.

ISO 8402, 1994. Quality management and quality assurance - Vocabulary, International Organization for Standardization (ISO).

ISO-TC/211, 2002. Geographic Information - Quality principles19113.

ISO-TC/211, 2003. Geographic Information - Quality evaluation procedures19114.

Juran J. M. Gryna F. M. J., Bingham R. S., *Quality Control Handbook*, New-York, McGraw-Hill, 1974.

Kahn B. K., Strong D. M., "Product and Service Performance Model for Information Quality: An Update." *Proceedings of Conference on Information Quality*, Cambridge, MA: Massachusetts Institute of Technology, 1998, p. 102-115.

Leitner M., Buttenfield B. P., "Guidelines for the Display of Attribute Certainty", *Cartography and Geographic Information Science*, vol. 27, nº 1, 2000, p. 3-14.

Longley P. A. Goodchild M. F. Maguire D. J., Rhind D. W., ed., 1999. *Geographical Information Systems*, John Wiley & Sons

Longley P. A. Goodchild M. F. Maguire D. J., Rhind D. W., ed., 2001. *Geographical Information Systems and Science*, John Wiley & Sons, 454 p.

Lowell K., "Why aren't we making better use of uncertainty information in decision-making?" *Proceedings of 6th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, Portland, Maine, USA, 2004.

Lowell K., Jaton A., *3rd International on Spatial Accuracy Assessment, Land Information Uncertainty in Natural Resources*, Quebec, Canada, Ann Arbor Press, 1999, 455 p.

Mac Eachren A. M., "Visualizing uncertain information", *Cartographic Perspectives*, vol. 13, 1992, p. 10-19.

Manche Y., *Analyse spatiale et mise en place de systèmes d'information pour l'évaluation de la vulnérabilité des territoires de montagne face aux risques naturels*, Thèse de doctorat, Université Joseph Fourier, Grenoble, 2000.

Martinet B., Marti Y.-M., *L'intelligence économique*, Éditions d'Organisation, 2001.

McGranaghan M., "A cartographic View of Spatial Data Quality", *Cartographica*, vol. 30, nº 2-3, 1993, p. 8-19.

Morrison J. L., 1974: "A theoretical framework for cartographic generalisation with the emphasis on the process of symbolisation". *International Yearbook of Cartography*, vol. 14, p. 115-127.

Morrison J. L., "Spatial data quality". *Elements of spatial data quality* (S. C. Guptill, and J. L. Morrison, Eds), Elsevier Science inc., New York, 1995.

Mowrer H. T., "Accuracy (Re)assurance: Selling Uncertainty Assessment to the Uncertain". *Spatial Accuracy Assessment, Land Information Uncertainty in Natural Ressources* (K. Lowell, and A. Jaton, Eds), Quebec, Ann Arbor Press, p. 3-10, 1999.

Office québécois de la langue française, 2004. www.olf.gouv.qc.ca

Paradis J., Beard K., "Visualization of Spatial Data Quality for the Decision Maker: A Data Quality Filter", *URISA Journal*, vol. 6, nº 2, 1994, p. 25-34.

Qiu J., Hunter G. J., "Managing Data Quality Information", *Proceedings of International Symposium on Spatial Data Quality*, Hong Kong, 18-20 July 1999, p. 384-395.

Qiu J., Hunter G. J., "A GIS with the Capacity for Managing Data Quality Information". *Spatial Data Quality* (W. Shi, M. F. Goodchild, and P. F. Fisher, Eds), Taylor & Francis, London, p. 230-250, 2002.

Reinke K. J., Hunter G. J., "A Theory for Communicating Uncertainty in Spatial Databases". *Spatial Data Quality* (W. Shi, P. F. Fisher, and M. F. Goodchild, Eds), Taylor & Francis, London, p. 77-101, 2002.

Schramm W., "How Communication Works". *Communication: Concepts and Processes* (J. A. DeVito, Eds), Prentice-Hall, New Jersey, p. 12-21, 1971.

Shannon C. E., "A Mathematical Theory of Communication", *The Bell System Technical Journal*, vol. 27, 1948, p. 379-423.

Simon H. A., "A Behavioral Model of Rational Choice?" *Quarterly Journal of Economics*, vol., nº 69, 1955, p. 99-118.

Sinton D. F., "The inherent structure of information as a constraint in analysis". *Harvard papers on Geographic Information Systems* (G. Dutton, Ed), Addison-Wesley, Reading, USA, 1978.

Smithson M., *Ignorance and Uncertainty: Emerging Paradigms*, New York, Springer Verlag, 1989.

Tastan H., Altan M. O., "Spatial Data Quality", *Proceedings of Third Turkish-German Joint Geodetic Days*, Istanbul, June 1-4, p. 15-30.

Taylor J. R., *An introduction to error analysis: the study of uncertainties in physical measurements*, Oxford, University Science Books, 1982.

Unwin D., "Geographical information systems and the problem of error and uncertainty", *Progress in Human Geography*, vol. 19, 1995, p. 549-558.

Veregin H., "Data quality parameters". *Geographical Information Systems* (P. A. Longley, M. F. Goodchild, D. J. Maguire, and D. W. Rhind, Eds), John Wiley & Sons, Inc., p. 177-189, 1999.

Wang R. Y., Strong D. M., "Beyond Accuracy: What Data Quality Means to Data Consumers", *Journal of Management Information Systems*, vol. 12, nº 4, 1996, p. 5-34.

Willett G., *La communication modélisée - Une introduction aux concepts, aux modèles et aux théories*, Éditions du Renouveau Pédagogique, Ottawa, 1992.

Windholz T. K., *Strategies for Handling Spatial Uncertainty due to Discretization*, Thèse de doctorat, University of Maine, Orono, 2001.

Chapitre 3 : Indicateurs de qualité

Indicateurs de qualité pour réduire les risques de mauvaise utilisation des données géospatiales.

R. Devillers, Y. Bédard et M. Gervais

Revue Internationale de Géomatique (2004), vol. 14, n. 1, p. 35-57

3.1 Résumé de l'article

Les utilisateurs de données géospatiales doivent être conscients de la qualité des données qu'ils manipulent afin de réduire les risques de mauvaises utilisations. L'information décrivant la qualité est variée et peut être représentée à différents niveaux de détails. Les utilisateurs peuvent donc accéder à de grands volumes d'information sur la qualité et se retrouver perdus dans cette abondance d'information. Cet article propose l'utilisation d'indicateurs de qualité pour améliorer la compréhension des informations relatives à la qualité des données géospatiales. Les concepts de tableau de bord et d'indicateur sont présentés et adaptés au domaine géospatial pour être intégrés dans des SIG. Un aperçu d'un prototype nommé Manuel à l'Usager Multidimensionnel (MUM) communiquant des indicateurs de qualité dans une interface de type SIG est présenté.

3.2 Introduction

Les domaines utilisant des données géospatiales sont variés et de nouvelles applications émergent fréquemment (Longley *et al.*, 1999). Si l'utilisation de données géospatiales était il y a quelques années un domaine réservé aux usagers experts utilisant des systèmes complexes et onéreux, la réalité a changé de manière significative (Hunter, 1999). Les logiciels SIG et les données géospatiales sont désormais accessibles à de faibles coûts, voire gratuitement sur Internet et sont de plus en plus faciles d'utilisation (Goodchild, 1995; Agumya et Hunter, 1997; Curry, 1998; Elshaw Thrall et Thrall, 1999). Les données géospatiales ne sont plus uniquement manipulées au niveau opérationnel dans les organisations mais également aux niveaux stratégique et tactique (Longley *et al.*, 1999). Elles sont maintenant de plus en plus utilisées pour supporter les processus de prise de décision (Hunter, 1999), allant de la sélection d'itinéraires pour planifier ses vacances, à la gestion d'un réseau routier par des agences gouvernementales. Avec le développement de services basés sur la localisation (LBS) et des technologies sans fils, il sera probablement habituel pour tout le monde dans un proche futur de prendre des décisions basées sur des données géospatiales visualisées sur des téléphones portables, systèmes nomades (ex. Palm Pilot), systèmes de navigation dans les voitures, etc.

Ces changements dans le contexte dans lequel les données géospatiales sont utilisées accroissent significativement les risques de mauvaises utilisations de ces données (Epstein *et al.*, 1998). Ainsi, Goodchild (1995) dit que « les SIG sont leurs propres ennemis : en invitant les gens à trouver de nouvelles utilisations des données, cela les invite aussi à être irresponsables dans leur utilisation » (traduction libre). Des cas de mauvaises utilisations sont fréquemment cités dans les revues scientifiques, les médias et les cas de jurisprudence (Blackmore, 1985; Beard, 1989; Goodchild et Kemp, 1990; Monmonier, 1994; Curry, 1998; Gervais, 2004). Beard (1989) identifie les mauvaises utilisations de données géospatiales comme étant des erreurs d'utilisation (*use error*), les ajoutant aux deux types d'erreur fréquemment mentionnés : les erreurs d'acquisition et les erreurs de traitement (*source errors* et *process errors).* Cette problématique favorise l'émergence de travaux visant à offrir des SIG pouvant prendre en compte la qualité des données manipulées (*quality-aware* ou *error-aware GIS*) (Buttenfield, 1993; Duckham et McCreadie, 1999;

Faïz, 1999; Hunter et Reinke, 2000; Duckham et McCreadie, 2002; Gan et Shi, 2002; Qiu et Hunter, 2002; Reinke et Hunter, 2002). Le problème est approché de différentes manières souvent complémentaires les unes des autres. Hunter et Reinke (2000) proposent de fournir des avertissements aux utilisateurs de SIG lorsque ceux-ci effectuent des « opérations illogiques ». Cette approche utilisant une base de règles s'adresse aux problèmes résultant de la manipulation de données avec les fonctions des SIG. Cette approche est de notre point de vue nécessaire mais n'apporte qu'une solution partielle au problème. En effet, beaucoup d'utilisateurs emploient les SIG à des fins de visualisation, sans utiliser d'opérateurs d'analyse (Lardon *et al.*, 2001; Roche, 2001). L'utilisateur du SIG peut ainsi effectuer des analyses de façon cognitive, comme identifier visuellement le chemin à suivre entre deux points de la carte, localiser la borne-fontaine la plus proche d'un bâtiment ou encore compter le nombre de bâtiments bordant une certaine rivière. Dans ces cas, des messages d'avertissement communiqués automatiquement ne pourraient pas prévenir les mauvaises utilisations. D'autre part, la base de règles nécessaire serait difficilement exhaustive, cohérente, et adaptée aux différents profils des utilisateurs, certaines règles pouvant par exemple être contextuelles à l'expertise ou au domaine d'application des utilisateurs. Cette approche ne peut donc pas prendre en compte tous les types de mauvaises utilisations. Différentes solutions peuvent alors être explorées pour réduire les risques de mauvaises utilisations des données géospatiales. Par exemple, Krek et Frank (1999) recommandent la création de jeux de données créés spécifiquement pour certains types d'utilisation (ex. navigation pédestre dans des environnements urbains), validant ainsi dès la production du jeu de données l'adéquation de celles-ci à leur utilisation. Si cette approche existe depuis longtemps (ex. les données topographiques étaient initialement produites pour des applications militaires), le contexte a changé. Les cartes de base (ex. topographique, cadastrales) sont à présent souvent utilisées à des fins différentes de leur objectif initial. D'un autre côté, basé sur une exploration des considérations légales reliées à l'utilisation de systèmes d'information géographiques (SIG), Gervais (2004) recommande, entre autres, aux utilisateurs novices de recourir à des experts en géomatique (*geomatics officer*) qui identifieraient les risques potentiels de mauvaises utilisations et confirmeraient, ou infirmeraient, l'adéquation de certains jeux de données à certaines utilisations. Une autre manière pouvant permettre la réduction des

risques de mauvaises utilisations, présentée dans cet article, consiste à fournir au sein de l'interface du SIG des informations contextuelles et compréhensibles sur la qualité des jeux de données manipulés. L'utilisateur étant informé de la qualité, si celle-ci semble problématique, il a alors le choix entre rechercher des données répondant mieux à ses besoins (réduisant ainsi l'incertitude) ou utiliser les données en étant conscient des conséquences éventuelles (absorbant ainsi l'incertitude résiduelle).

Les métadonnées (*i.e.* données sur les données) distribuées par certains producteurs de données fournissent déjà une partie de ces informations. Toutefois, les métadonnées actuellement fournies sont plus des descriptions techniques des jeux de données que des informations compréhensibles destinées aux usagers (Timpf *et al.*, 1996; Harvey, 1998). Les métadonnées sont donc dans la pratique très peu utilisées, laissant les utilisateurs experts comme non-experts en géomatique dans un état d'ignorance concernant la qualité des données géospatiales qu'ils manipulent.

De plus, les métadonnées sont très rarement assez détaillées (Hunter, 2001; Gan et Shi, 2002), étant la plupart du temps une description au niveau du jeu de données. La qualité décrite dans les métadonnées est alors une agrégation de qualités hétérogènes des objets composant le jeu de données (ex. « la précision spatiale du jeu de données varie entre 10 m et 1 km »). Pour être utiles, celles-ci devraient décrire les données à un niveau de détail plus fin, comme au niveau de l'instance d'objet ou même de la valeur d'un attribut, ce qui fournirait un grand volume d'information aux utilisateurs. Toutefois, les humains ne résolvent pas les situations complexes avec un grand volume d'information, mais en sélectionnant les informations pertinentes (Klein, 1999). Dans ce sens, Fisher (2001) mentionne que « le défi dans un monde riche en information n'est pas seulement de rendre l'information disponible aux personnes en toute place et sous toutes les formes, mais de réduire la surcharge d'information en rendant l'information pertinente pour la tâche effectuée et en fonction du bagage de connaissances présumé des utilisateurs » (traduction libre).

Ainsi, afin de fournir uniquement des informations pertinentes aux usagers au niveau de détail dont ils ont besoin, il existe un besoin pour différentes vues agrégées de ces

informations de qualité, contextuelles aux profils des utilisateurs et à la tâche qu'ils effectuent.

Ce besoin n'est pas exclusif aux données géospatiales. Par exemple, les gestionnaires d'entreprises ont également besoin d'une vue agrégée des informations décrivant leur compagnie. Les gestionnaires de grandes chaînes de magasins ne sont en général pas intéressés par des listes de toutes les ventes faites dans leurs magasins, mais par des agrégations des ventes par type de produits, intervalle de temps, région, etc. Dans ces domaines, les décideurs utilisent maintenant des outils du domaine du *Business Intelligence* les aidant dans leurs tâches. Parmi ces outils, les tableaux de bord exécutifs fournissent des informations agrégées, nommées indicateurs, sur différents aspects des organisations.

L'objectif de cet article est de présenter une approche basée sur des indicateurs permettant de communiquer l'information relative à la qualité des données géospatiales aux utilisateurs. Étant donné que les problèmes de qualité deviennent rapidement complexes (ex. diversité des paramètres décrivant la qualité, granularité des informations sur la qualité, hétérogénéité spatiale), les utilisateurs ciblés dans cet article sont des experts ayant une bonne connaissance de la géomatique. En effet, dans le contexte actuel, même les experts en géomatique ont beaucoup de difficulté à se prononcer sur la qualité des données pour une application précise dans un secteur précis. Les données utilisées résultent souvent de l'intégration de différentes sources pouvant avoir été collectées suivant différentes normes, à différentes époques avec des technologies diverses. Ainsi, les données manipulées dans les SIG sont souvent très hétérogènes et l'adéquation de ces données à une application spécifique demeure complexe. Bien que l'application présentée vise des usagers experts en géomatique, l'approche globale est en partie applicable à des utilisateurs non-experts. Cette approche fournit des informations pertinentes aux utilisateurs relativement à la qualité des données qu'ils manipulent afin de réduire les risques de mauvaises utilisations de ces données. L'utilisation de logiciels SIG dans des processus de prise de décision sera abordée dans la section 3.3. De manière plus spécifique, nous présenterons les SIG comme un processus de communication entre des producteurs et des utilisateurs de données. Nous présenterons l'incertitude reliée aux processus de prise de décision et mentionnerons les informations relatives à la qualité des données actuellement communiquées aux utilisateurs et leurs limites pour le support à la prise de décision. La section 3.4 présente les concepts et

caractéristiques des tableaux de bord de gestion et des indicateurs. Les caractéristiques des tableaux de bord et indicateurs dans le contexte géospatial seront présentées. Finalement, la section 3.5 présente un aperçu d'un prototype intégré dans une interface cartographique, permettant de gérer et communiquer ces indicateurs. Le prototype sera présenté plus en détails dans le chapitre 5.

3.3 SIG et prise de décision

3.3.1 SIG – Un processus de communication

Shannon (1948) définit la communication comme « reproduire en un point exactement ou approximativement un message sélectionné en un autre point » (traduction libre). Basé sur les adaptations de la théorie de la communication de Shannon pour le domaine de la communication de masse (ex. journalisme) et pour les sciences cognitives (ex. perception, interprétation de signaux), Bédard (1987) identifie les SIG (en tant que système organisationnel) comme étant un processus de communication complexe entre des producteurs et des utilisateurs de données géospatiales. Afin de prendre une décision, les personnes perçoivent des signaux du monde réel, les interprètent, et procèdent à une abstraction afin de générer un modèle cognitif servant à cette prise de décision. Les signaux perçus peuvent provenir soit d'une observation directe de la réalité, soit d'une autre personne (ou machine) mandatée pour communiquer une information. Dans le cas des utilisateurs de logiciels SIG, les signaux perçus proviennent presque toujours d'un observateur autre que l'utilisateur, créant ainsi un processus de communication entre l'observateur de la réalité (ex. géomètre, forestier, géologue) et l'utilisateur du logiciel SIG. De nos jours, il est même de plus en plus fréquent pour un utilisateur de logiciel SIG d'utiliser des données multisources.

Une caractéristique importante des processus de communication est le besoin de connaissances communes (identifié en anglais par le concept de *commonness*) entre producteurs de signaux et récepteurs (pouvant être des individus ou des machines) (Shannon, 1948; Bédard, 1987; Martinet et Marti, 2001). L'ensemble des connaissances d'un agent est identifié comme étant son cadre de référence. Plus les *connaissances communes* sont importantes entre le producteur et l'utilisateur d'une information, plus les

risques de distorsion du message sont faibles. En pratique, cette communication est toujours imparfaite à cause des différences entre sources et cibles. Afin de faciliter la communication entre les agents, Martinet et Marti préconisent l'utilisation d'un langage le plus proche possible de la cible. Les SIG communiquent donc toujours les informations avec un certain biais, mais l'emploi d'un langage graphique proche des connaissances des utilisateurs des données peut limiter ce biais.

3.3.2 Prise de décision et incertitude

Les utilisateurs de SIG manipulent les données géospatiales afin d'obtenir des informations pouvant être utilisées dans un processus de prise de décision plus large (ex. prendre le chemin le plus court pour se rendre quelque part, trouver la parcelle cadastrale idéale pour construire un bâtiment). Mintzberg (1979) définit la décision comme « le signal d'une intention explicite d'agir » (traduction libre). La décision ne se limite pas à l'action. Fernandez (2000) identifie quatre étapes dans un processus de prise de décision, soit (1) la formalisation du désir, lorsque l'agent prend conscience de la situation, (2) l'instruction, lorsque l'agent collecte les informations, analyse des situations précédentes et des solutions potentielles, (3) le choix, lorsque l'agent identifie l'action à effectuer et évalue ses limites et enfin (4) l'action. Il formule également plusieurs conditions pour prendre une bonne décision. Les décisions sont prises (1) afin d'atteindre un objectif, (2) selon la situation perçue, (3) selon l'expérience et le référentiel de valeurs du décideur, (4) selon ses motivations, (5) en fonction de la mesure des risques et (6) selon les moyens conférés et disponibles. Basé sur des observations pratiques de différents types de décideurs, Klein (1999) affirme que l'intuition et les simulations mentales sont centrales dans la prise de décision, basées respectivement sur l'expérience et l'imagination. Il explique que « l'intuition dépend de l'usage de l'expérience pour reconnaître des patrons clés indiquant la dynamique de la situation » (Klein, 1999; p. 31 – traduction libre). Le modèle RPD de Klein (*Recognition-Primed Decision*) offre un cadre théorique pour les processus de prise de décision. Ce modèle souligne l'importance des indices pertinents qui aident les décideurs à reconnaître une situation, évitant une surcharge possible d'information.

Fernandez (2000) différencie la décision du calcul. D'un côté, le calcul permet le choix d'une bonne solution rationnelle et est automatisable. De l'autre, la décision est basée sur

des informations incertaines, imprécises et insuffisantes, mettant en jeu le contexte, les acteurs et la situation. Cette incertitude peut apparaître à différents niveaux. Le concept d'incertitude, ainsi que d'autres termes liés au domaine de la qualité, est présenté, entre autres, plus précisément dans un article de Fisher (1999).

Quand une personne fait face à de l'incertitude lors d'un processus de prise de décision et est consciente du type d'incertitude et de son importance, il peut choisir entre (1) ne rien faire, (2) essayer de réduire cette incertitude ou (3) prendre la décision et accepter les conséquences possibles, « absorbant » ainsi cette incertitude (Bédard, 1987). Epstein *et al.* (1998) suggèrent de réduire l'incertitude en (1) obtenant plus d'information et/ou (2) améliorant la qualité de l'information disponible. L'incertitude résiduelle alors absorbée est alors à la source du risque relié à l'utilisation de cette information (Bédard, 1987; Epstein *et al*., 1998). Le niveau de risque acceptable dépend du décideur, de l'application ou du contexte institutionnel.

Les décisions sont donc toujours basées sur des informations incertaines et incomplètes. Les décideurs ont alors le choix entre prendre la décision en acceptant l'incertitude résiduelle ou collecter de nouvelles informations pour diminuer cette incertitude. Ceux-ci utilisent des indices (ou indicateurs) afin de caractériser une situation, diminuer l'incertitude et donc orienter leur décision.

3.3.3 Communication de l'information sur la qualité des données géospatiales

Les producteurs de données fournissent de plus en plus souvent des métadonnées documentant différents aspects des jeux de données, afin de renseigner les utilisateurs sur les caractéristiques des données qu'ils utilisent. Selon les principales normes en géomatique, les métadonnées devraient fournir de l'information relative à la qualité des données géospatiales, telle que la précision spatiale, la complétude (omission, commission) ou la consistance logique de la base de données (Guptill et Morrison, 1995; FGDC, 2000; ISO-TC/211, 2003).

Toutefois, l'expérience montre que ces métadonnées sont complexes à comprendre et à utiliser pour des utilisateurs non-experts mais aussi par les experts en données géospatiales, restant de ce fait la plupart du temps inutilisées (Timpf *et al*., 1996; Frank, 1998). Ce sont

plus des descriptions techniques dont le contenu découle des procédures de production des jeux de données, que des informations compréhensibles et pertinentes pouvant être utilisées par des utilisateurs de données pour supporter leur processus de prise de décision (Frank, 1998).

De plus, les métadonnées fournissent la plupart du temps une description des données au niveau du jeu de données. Comme la qualité peut être très hétérogène dans l'espace et dans le temps, des métadonnées à un niveau de détail plus fin, tel que l'occurrence d'objet ou l'attribut, seraient souvent nécessaires (Hunter, 2001; Gan et Shi, 2002).

Les métadonnées ne sont donc pas un moyen efficace de communiquer les informations relatives à la qualité des données aux utilisateurs de données. Différentes approches peuvent alors être explorées pour aborder ce problème.

Basé sur le paradigme de communication des SIG développé par Bédard, la Figure 10 illustre de façon théorique quelques solutions possibles, en décrivant les relations entre les connaissances des producteurs et utilisateurs de données ainsi que la position des métadonnées dans ce cadre de référence. Les cercles représentent les cadres de référence des producteurs et des utilisateurs de données et leurs intersections correspondent aux connaissances communes.

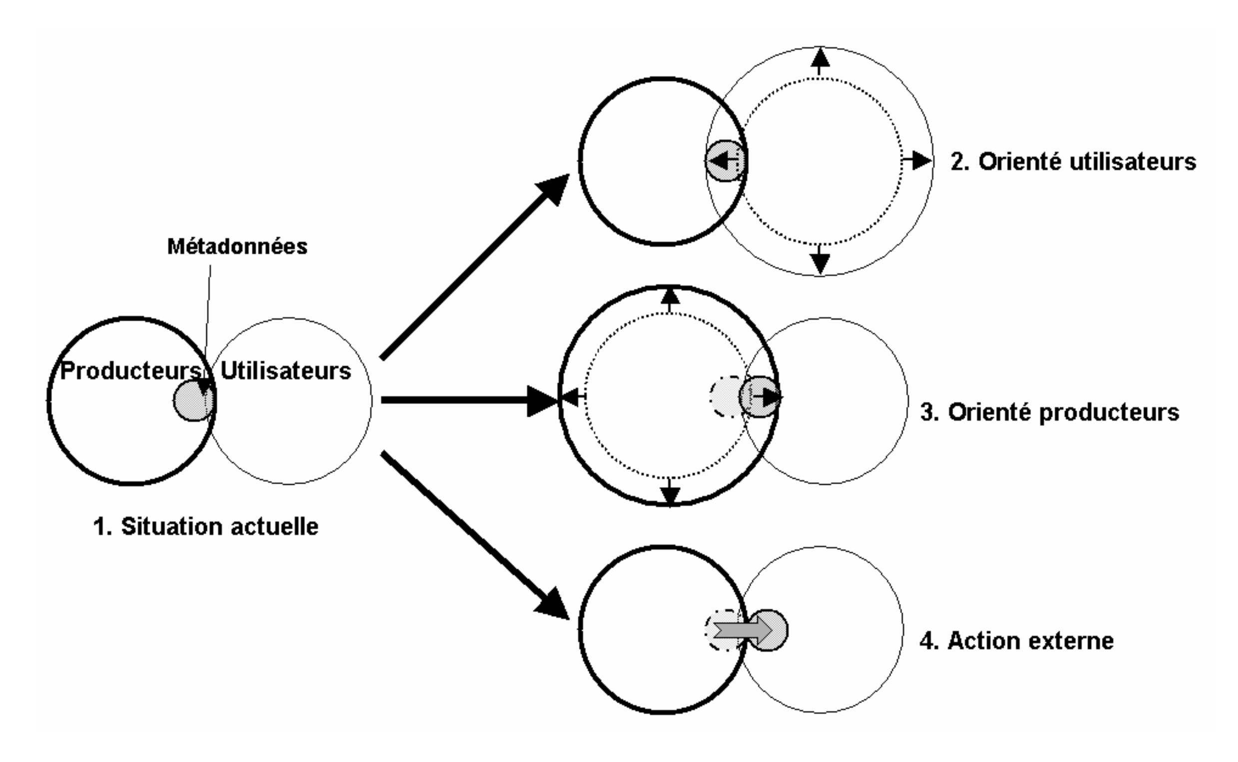


Figure 10 : Les métadonnées dans le processus de communication utilisateurs-producteurs.

1. Situation actuelle : Actuellement, les producteurs de données et les utilisateurs ont des connaissances communes faibles concernant les données géospatiales. Les métadonnées utilisent un vocabulaire technique, et sont donc situées dans le cadre de référence des producteurs de données. Les utilisateurs n'ayant généralement pas de connaissance dans le domaine de la géomatique (projections, échelles, techniques d'acquisition et de traitement des données, etc.), ils ont une compréhension très réduite des métadonnées.

2. Approche orientée utilisateurs : Une solution possible serait d'accroître le cadre de référence des utilisateurs, en leur « enseignant » des concepts et du vocabulaire relié aux données géospatiales. Ceci augmenterait la connaissance des utilisateurs, accroissant leur expertise et donc les connaissances communes entre producteurs et utilisateurs. Toutefois, cela signifie faire devenir experts des utilisateurs non-experts.

3. Approche orientée producteurs : L'inverse serait d'accroître le cadre de référence des producteurs – *e.g.* essayer de vulgariser le vocabulaire technique des métadonnées– accroissant ainsi les connaissances communes. Certaines normes recommandent une telle approche en encourageant l'utilisation de vocabulaire non technique et les descriptions en

texte libre afin de rendre plus accessible la compréhension des métadonnées. Toutefois, cela représente un travail additionnel de la part des producteurs de données et demeure souvent insatisfaisant en regard de la compréhension des utilisateurs.

Ce problème est semblable à des problèmes de communication classiques entre deux individus ne parlant pas la langue de l'autre (ex. un chinois et un français). Les cas 2 et 3 signifient qu'une des deux personnes apprend le langage de l'autre. Toutefois, comme cela requière des ressources considérables (temps et souvent argent), il est fréquent de faire appel à une troisième personne connaissant les deux langages pour traduire le message.

4. Action externe : Ce cas, en linguistique, correspond au rôle du traducteur qui est chargé de faire le lien entre les deux agents. Le message produit par l'un des agents (*i.e.* producteur de données) dans un langage (*i.e.* métadonnées techniques) est traduit dans le langage de l'autre agent (*i.e.* utilisateur des données). Comme tout processus de communication, la distorsion du message est minimisée mais souvent inévitable. Ce dernier cas est l'approche examinée dans cet article, traduisant des métadonnées stockées en général dans des fichiers texte externes aux données, en des indicateurs plus facilement interprétables intégrés dans l'interface du SIG. L'approche présentée va également plus loin en effectuant cette traduction de manière contextuelle, en offrant aux usagers un accès hiérarchique aux indicateurs et en l'avertissant potentiellement de risques de mauvaises utilisations.

3.4 Tableaux de bord et indicateurs pour supporter la prise de décision

3.4.1 Tableaux de bord

L'analogie avec les tableaux de bord automobiles peut illustrer la façon de représenter une réalité complexe en utilisant un modèle simplifié. Le tableau de bord d'une voiture permet au conducteur d'obtenir en temps réel des informations limitées, mais souvent suffisantes, concernant le système plus complexe qu'est son véhicule. Même si le tableau de bord donne une vision incomplète et souvent imprécise de la réalité, cela permet au conducteur de prendre des décisions rapides telles que diminuer sa vitesse, ajouter de l'huile ou s'arrêter prendre de l'essence.

Un tableau de bord de gestion est défini par Voyer comme « une façon de sélectionner, d'agencer et de représenter des indicateurs essentiels et pertinents, de façon sommaire et ciblée […] fournissant à la fois une vision globale et la possibilité de forer dans les niveaux de détail. » (Voyer, 2000; p.39). Dans le domaine des systèmes de support à la prise de décision, les tableaux de bord de décision (*executive dashboard*) sont aussi nommés *scorecard, balanced scorecard, scoreboard, steering panel* ou *control panel*. Les tableaux de bord se concentrent surtout sur la qualité de l'information et non sur sa quantité. Ils représentent les indicateurs de façon compréhensible, suggestive et attractive afin de faciliter leur visualisation. Ils présentent un aperçu représentatif de la situation, permettant ensuite d'accéder aux données plus détaillées au besoin. Le tableau de bord doit être contextuel, le décideur pouvant sélectionner ses propres indicateurs, avec la représentation qu'il préfère, afin de produire son tableau de bord personnalisé.

De nombreuses organisations utilisent des tableaux de bord. Par exemple, le gouvernement canadien encourage l'utilisation d'indicateurs dans son administration. Les grandes compagnies comme les banques et les compagnies d'assurances utilisent des indicateurs. Les grands organismes internationaux (Banque Mondiale, Nations Unies, agences américaine et canadienne de développement international, etc.) utilisent également des indicateurs sociaux, économiques, géopolitiques ou environnementaux.

Les tableaux de bord permettent la visualisation d'un ensemble d'indicateurs. En effet, l'utilisation d'un seul indicateur serait trop dangereuse (Kaplan et Norton, 1992). Prenez par exemple un pilote d'avion qui a besoin d'information sur de nombreuses variables telles que l'essence, l'altitude, la vitesse de l'air, la position, la destination, etc. Ces informations ne peuvent pas être fournies par un seul indicateur. Le nombre d'indicateurs doit cependant être limité afin d'éviter une surcharge d'information. D'après Miller (1956), l'être humain peut percevoir 7±2 éléments en même temps. Ce « nombre magique » est maintenant largement utilisé pour la communication d'informations et peut donc être utilisé dans la conception de tableaux de bord en géomatique.

3.4.2 Indicateurs

Le *Jackson Community Council* (Plan Canada, 1999) définit un indicateur comme « une manière de voir un portrait général en regardant un petit morceau de celui-ci » (traduction

libre). Fernandez (2000) le définit comme « une information ou un regroupement d'informations contribuant à l'appréciation générale d'une situation par le décideur » (p.232). Klein (1999) identifie les indicateurs comme étant des indices situés au centre des processus de prise de décision en supportant les intuitions des décideurs. L'objectif d'un indicateur est de mesurer une situation et d'initier une réaction du décideur, la réaction pouvant être de ne rien faire.

Le système doit fournir un ensemble d'indicateurs que les usagers peuvent adapter à leurs contextes si besoin ou des indicateurs pouvant être partagés à l'intérieur d'une même communauté d'utilisateurs. Il devrait aussi permettre aux décideurs de créer leurs propres indicateurs et règles pour les calculer.

La valeur d'un indicateur peut être basée sur une donnée unique ou résulter d'un calcul impliquant plusieurs données. Ces données doivent être techniquement accessibles. Elles peuvent être déjà disponibles dans une base de données accessible ou provenir d'autres sources, telles que des opinions d'experts ou de collègues. Comme les données sont valides pour une certaine durée dans le temps (*life time*), leur actualité doit être prise en considération.

Les caractéristiques des indicateurs peuvent être décrites sur une feuille d'indicateurs que les utilisateurs peuvent consulter et modifier si nécessaire. Cette feuille peut fournir par exemple de l'information sur la définition de l'indicateur, ses représentations possibles, sur les considérations reliées à son utilisation et interprétation, ses mécanismes de validation, etc. Les indicateurs peuvent représenter différents types d'information, tant quantitatifs que qualitatifs. Il est préférable de fournir des indicateurs « flous », *i.e* de précision limitée (Fernandez, 2000), tels qu'un intervalle de valeurs ou une échelle qualitative, car des valeurs trop précises encourageraient l'utilisateur à se concentrer sur la valeur et non sur sa signification dans une perspective globale.

Diverses représentations peuvent être utilisées pour visualiser la valeur d'un indicateur, telles que des nombres, symboles, icônes, pictogrammes, tables, graphiques, textes, sons, images, etc. Il est également possible d'utiliser des fenêtres *pop-up*, alarmes visuelles ou sonores, etc., qui sont souvent des façons efficaces de capter l'attention des utilisateurs afin qu'ils se concentrent sur l'essentiel.

3.5 Tableaux de bord et indicateurs pour la prise de décision géospatiale

3.5.1 Tableaux de bord et système MUM

Les tableaux de bord de gestion se rattachent au domaine du support à la prise de décision et de manière plus spécifique du *Business Intelligence*. Certains travaux ont été faits pour adapter des outils du *Business Intelligence* dans le domaine de la géomatique, tels que pour le *Data Mining* Spatial, le SOLAP (*Spatial On-Line Analytical Processing*) et les entrepôts de données géospatiales (Miller et Han, 2001; Rivest *et al.*, 2001). Plusieurs logiciels développés pour le domaine du *Business Intelligence* visent à créer et maintenir des tableaux de bord de gestion. Ils sont par exemple *Esperant* et *Media* de Speedware, *Metrics Manager* de Cognos, *EIS* de SAS, *Oracle Balanced Scorecard* de Oracle, *Hyperion Performance Scorecard* de Hyperion, *Crystal Application* de Crystal Decisions. Comme les tableaux de bord fournissent habituellement des informations à différents niveaux de détails, la plupart des systèmes reposent sur des bases de données multidimensionnelles. Une telle structure conçue pour la gestion des informations relatives à la qualité des données géospatiales est décrite dans le chapitre 4 et est utilisée pour la conception du prototype de tableau de bord géospatial. Cette structure permet de gérer les informations de qualité à différents niveaux de détails.

Les fonctionnalités du tableau de bord devraient s'inspirer des concepts énoncés dans les sections précédentes, tel que communiquer des informations sur une base visuelle, éviter une surcharge d'information, permettre aux utilisateurs d'adapter leur tableau de bord à leurs besoins (ex. choix des indicateurs, type de visualisation, type de calcul des indicateurs), etc. En plus des fonctionnalités offertes par un tableau de bord « classique », la composante spatiale doit elle aussi être prise en compte. Le tableau de bord de qualité devrait donc être capable de :

- *Représenter l'information de qualité sous la forme d'indicateurs* : les indicateurs fournissent des informations brutes ou agrégées sur la qualité des données géospatiales. Les indicateurs doivent être présentés sur un tableau de bord faisant partie de l'interface du SIG et peuvent être rendus visibles ou non selon le désir des utilisateurs;

- *Fournir des indicateurs en temps réel :* étant donné que les utilisateurs peuvent vouloir ajouter ou retirer des données dans leurs SIG, modifier leurs profils personnels (ex. tolérance face au risque), etc., les valeurs des indicateurs doivent être recalculées à chaque modification du contexte de l'utilisateur.

- *Fournir des indicateurs en fonction de l'étendue spatiale visualisée* : la qualité peut être très hétérogène dans l'espace et dans le temps. Par exemple, un secteur d'une carte a pu être mis à jour récemment avec une grande précision et exactitude tandis qu'un autre secteur de la même carte présente des données anciennes et imprécises. Les valeurs des indicateurs doivent donc être calculées à partir des qualités des objets situés dans la zone visualisée par l'utilisateur, et non pas uniquement représenter la qualité moyenne de l'ensemble des données du jeu de données. L'utilisateur doit également pouvoir obtenir la qualité moyenne d'une zone qu'il définit de façon ad hoc, soit en lui permettant de tracer lui-même cette zone (ex. création d'un polygone), soit en lui offrant une liste de zones prédéfinies (ex. villes, quartiers). Cela implique une mise à jour des indicateurs lorsque l'utilisateur navigue dans sa vue (ex. *Zoom in, Zoom out*, *Pan*);

- *Permettre aux utilisateurs de sélectionner les indicateurs pertinents dans leur contexte ou définir leurs propres indicateurs* : différents utilisateurs ont des profils, objectifs et intérêts différents. Un ensemble d'indicateurs prédéfinis doit être mis à la disposition des utilisateurs. Toutefois, les utilisateurs doivent être capables de voir comment sont calculés ces indicateurs, de modifier ces procédures et si possible de permettre la création de nouveaux indicateurs;

- *Permettre aux utilisateurs de visualiser les indicateurs à différents niveaux de détails* : les indicateurs doivent être organisés de manière hiérarchique (indicateurs et sous-indicateurs) afin d'éviter aux utilisateurs une surcharge d'information. On conserve ainsi un nombre d'indicateurs conforme à la loi de Miller (7±2 indicateurs) tout en permettant aux utilisateurs d'approfondir l'exploration des informations de qualité de manière intuitive;

- *Permettre aux utilisateurs de mettre des poids sur les différents indicateurs, en fonction de leur importance dans le contexte d'utilisation des données* : certains indicateurs peuvent avoir plus d'importance que d'autres. Par exemple, la complétude des données peut être beaucoup plus importante que la précision temporelle pour certaines applications. Ces

poids entrent en jeu lors de l'agrégation des sous-indicateurs en indicateurs de plus hauts niveaux ;

- *Permettre la définition et la gestion des profils des utilisateurs (niveau de risque acceptable, etc.)* : différents utilisateurs peuvent vouloir différentes façons d'agréger des indicateurs. Par exemple, certaines personnes peuvent avoir une plus grande tolérance face aux risques que d'autres dans leurs décisions, suivant par exemple leur contexte organisationnel (ex. une personne utilisant un SIG pour planifier une sortie de loisir en famille pourra accepter plus de risques qu'un gestionnaire utilisant un SIG pour gérer des épidémies dans un organisme de santé environnementale);

- *Offrir différentes représentations des indicateurs que les utilisateurs peuvent sélectionner* : certains utilisateurs peuvent préférer certaines représentations pour les indicateurs (ex. feux de circulation, histogrammes, compteur de vitesse). Les utilisateurs doivent pouvoir choisir le mode de représentation qu'ils préfèrent parmi un choix de représentations dépendant du type d'indicateur et des valeurs qu'il communique (ex. quantitatif, qualitatif);

- *Offrir un mode de visualisation cartographique des indicateurs de qualité* : en plus d'une représentation des indicateurs dans un tableau de bord, la valeur des indicateurs doit pouvoir être représentée sur la carte. Par exemple, un indicateur ayant une représentation du type feu de circulation (vert/jaune/rouge) pourra avoir une valeur jaune représentant l'ensemble des données visualisées dans l'interface. L'utilisateur pourra passer en mode de représentation cartographique de la qualité et ainsi avoir une meilleure idée des qualités individuelles des objets (ex. précision spatiale), chaque objet affiché étant représenté en vert, jaune ou rouge, dépendamment de sa qualité. Cette représentation permet entre autres d'identifier rapidement l'hétérogénéité spatiale de la qualité.

- *Activer des alarmes automatiquement lorsque certaines conditions sont atteintes* : des signaux sonores ou visuels peuvent être émis pour capter l'attention des utilisateurs à certains moments critiques, comme lorsqu'un indicateur dépasse la tolérance définie par l'utilisateur.

La Figure 11 présente un schéma général de la création des indicateurs qui seraient affichés dans le tableau de bord du système MUM : (1) Une interface permet de collecter les

informations caractérisant l'utilisateur (contexte, style de gestion, etc.) et conserve ces informations. (2) Une base de données d'indicateurs prédéfinis permet à l'utilisateur de sélectionner et éventuellement de modifier des indicateurs existants. L'utilisateur peut aussi définir et stocker de nouveaux indicateurs. Cette étape permet aux utilisateurs de personnaliser leurs indicateurs et leur tableau de bord. (3) Les métadonnées et autres informations pertinentes décrivant les jeux de données sont intégrées et structurées dans une même base de données à différents niveaux de détails. Ce processus d'intégration doit idéalement être automatique ou semi-automatique afin d'assurer une certaine flexibilité au système. (4) Les indicateurs ayant été sélectionnés, leurs valeurs sont calculées en utilisant la règle d'agrégation définie, celle-ci dépendant de l'indicateur, de l'information disponible pour le calculer et du profil de l'usager. (5) Les indicateurs sélectionnés sont alors affichés dans l'interface du SIG selon le mode de représentation choisi par l'usager afin d'informer l'utilisateur de la qualité des données qu'il utilise. Ces indicateurs sont par la suite mis à jour dès que des changements ont lieu (ex. changement au profil de l'utilisateur, navigation dans l'interface cartographique, navigation à l'aide de fonctions OLAP).

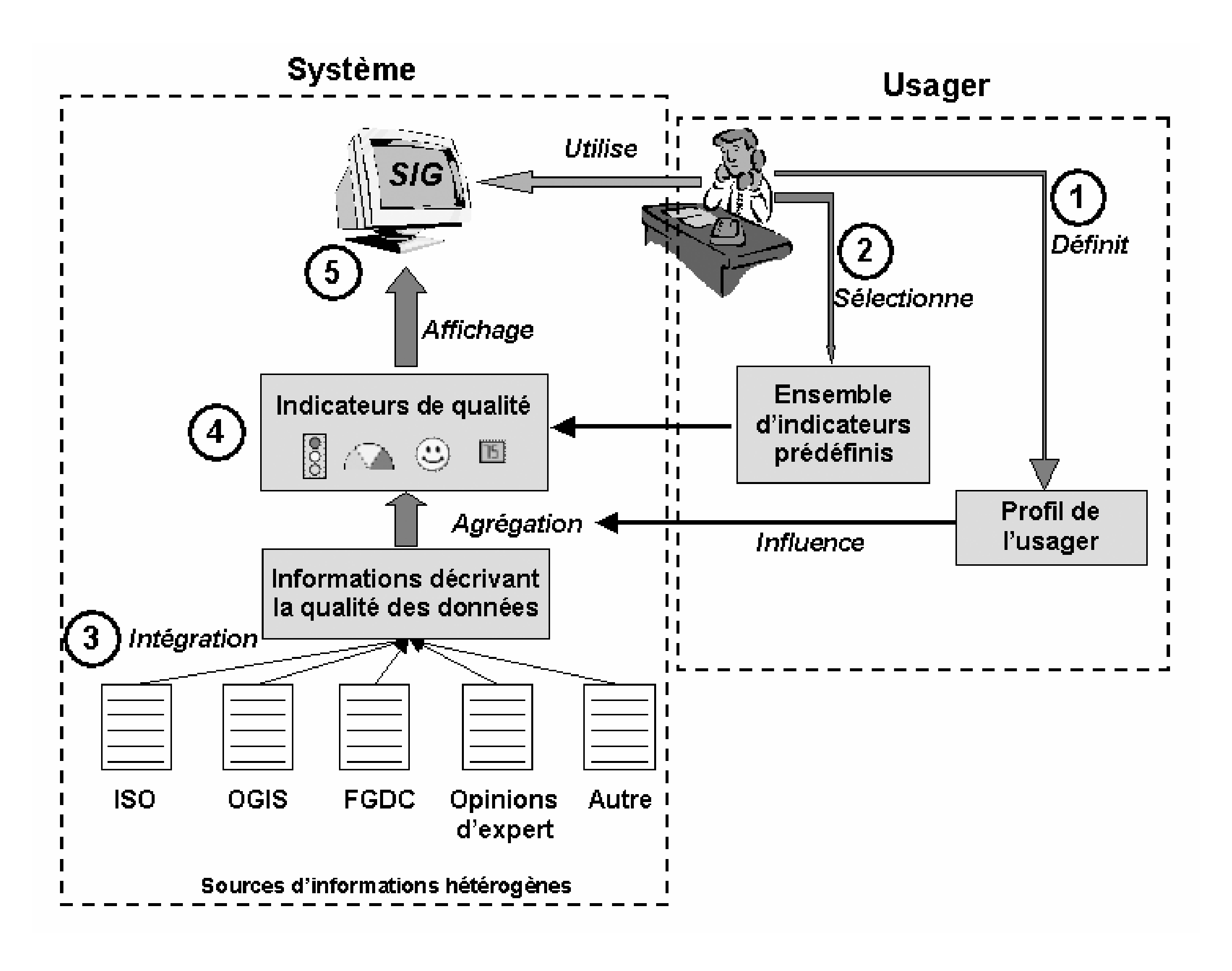


Figure 11 : Fonctionnement simplifié du système MUM.

3.5.2 Indicateurs de qualité des données géospatiales

Les mauvaises utilisations de données géospatiales peuvent apparaître à différents niveaux lors de l'utilisation d'un SIG. Par exemple, un utilisateur peut faire une mauvaise utilisation d'une fonction d'un SIG (ex. interpoler des données nominales de points ou afficher au 1 :10 000 une carte créée à l'échelle 1 :1 000 000). Un utilisateur peut également obtenir un mauvais résultat de fonctions dans un SIG lorsque les données contiennent des erreurs (ex. mesures de distances très précises basées sur des données largement inexactes ou encore calculer un nombre d'objets alors que la complétude du jeu de données est médiocre). Il est également nécessaire de prendre en considération que beaucoup d'utilisateurs font appel aux SIG pour visualiser les données, sans forcément utiliser de fonctions d'analyse. Donc, les outils visant à réduire les risques de mauvaises utilisations de données géospatiales

doivent se concentrer à la fois sur les erreurs issues de la manipulation des opérateurs d'un SIG (ex. opérateurs topologiques et métriques) et sur la mauvaise interprétation de données affichées par le SIG. Nous pouvons ainsi identifier deux types d'avertissements pouvant réduire les risques de mauvaises utilisations :

— Avertissements de manipulation :

- Messages d'opérations illogiques (Hunter et Reinke, 2000) : des avertissements sonores ou visuels peuvent être communiqués aux utilisateurs lorsqu'une manipulation pouvant engendrer un risque est effectuée sur des données dans le SIG (ex. requêtes, zoom, mise à jour) (cf. Figure 12). Hunter et Reinke donnent plusieurs exemples d'opérations illogiques pouvant être traduites en algorithmes tels que :

```
IF command_name = 'calculate_map_distance'
     AND map_units = null
     OR distance_units = null
     OR projection_type = null
THEN generate map_distance_warning
```

De telles règles pourraient limiter les risques les plus courants de manipulation en émettant des avertissements ou en désactivant certaines fonctions du SIG pouvant induire un risque. Une connaissance des données, provenant par exemple des métadonnées (ex. exactitude des données) ou directement de la structure des données (ex. précision numérique des données), est nécessaire. Les règles doivent être définies par des experts puis stockées dans une base de règles pouvant être interrogée par le système lors de chaque opération effectuée dans le SIG.

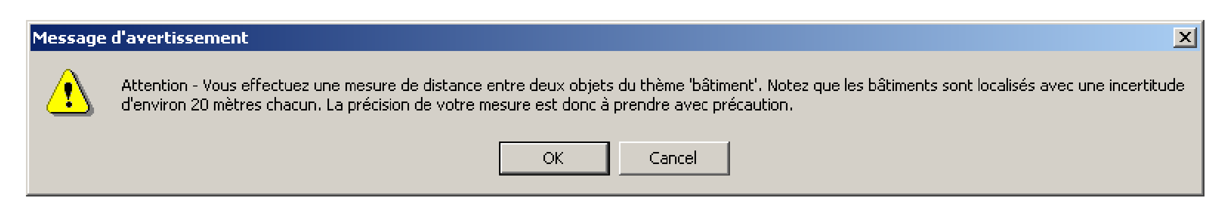


Figure 12 : Exemple de message d'opération illogique.

— Avertissements d'état :

- Indicateurs de statut : un indicateur de statut fournit une information « brute » décrivant une ou plusieurs caractéristiques internes des données. Par exemple, si la précision spatiale est de 13 mètres, la valeur de l'indicateur « précision spatiale » pourrait être par exemple « 13 mètres » ou un intervalle tel que « 10-20 mètres »;

- Indicateurs de risque : les indicateurs de risque fournissent une information « relative », provenant de la comparaison entre des caractéristiques des données (données brutes) et les besoins des utilisateurs, la divergence étant exprimée en terme de niveau de risque. Ceux-ci représentent des informations sur la qualité des données, la qualité étant définie comme l'adéquation à l'usage (*fitness for use*). Par exemple, si la précision spatiale des données est de « 13 mètres » et l'usager désire des données avec une précision de « 1 mètre », la valeur de l'indicateur précision spatiale pourrait être affichée avec une lumière rouge, avertissant l'utilisateur de l'importance de la différence. L'utilisateur aurait alors une idée du risque relié à l'utilisation des données pour ce critère de qualité. Le calcul de ce type d'indicateur implique une qualification de données quantitatives (passer de « 13 mètres » à une lumière rouge dans cet exemple). Cette qualification de l'information est complexe et peut être faite de différentes façons. Différents modes de représentation peuvent être utilisés pour les indicateurs de risque, tels que des feux de circulation, des compteurs de vitesse, des *smiley*, etc. Ces indicateurs généralement binaires ou ternaires permettent de représenter un indicateur passant un message du type *go/no go* ou mauvais/moyen/bon.

Les utilisateurs doivent avoir accès à des descriptions des indicateurs proposés. Un exemple de fiche descriptive d'indicateurs est présenté sur la Figure 13. Cette fiche permet la description de différents aspects de l'indicateur tels que :

- Définition/signification de l'indicateur;

- Méthode utilisée pour calculer la valeur de l'indicateur;

- Mode de représentation (ex. valeur simple, feux de circulation, *smiley*);

- Importance de l'indicateur pour l'utilisateur;

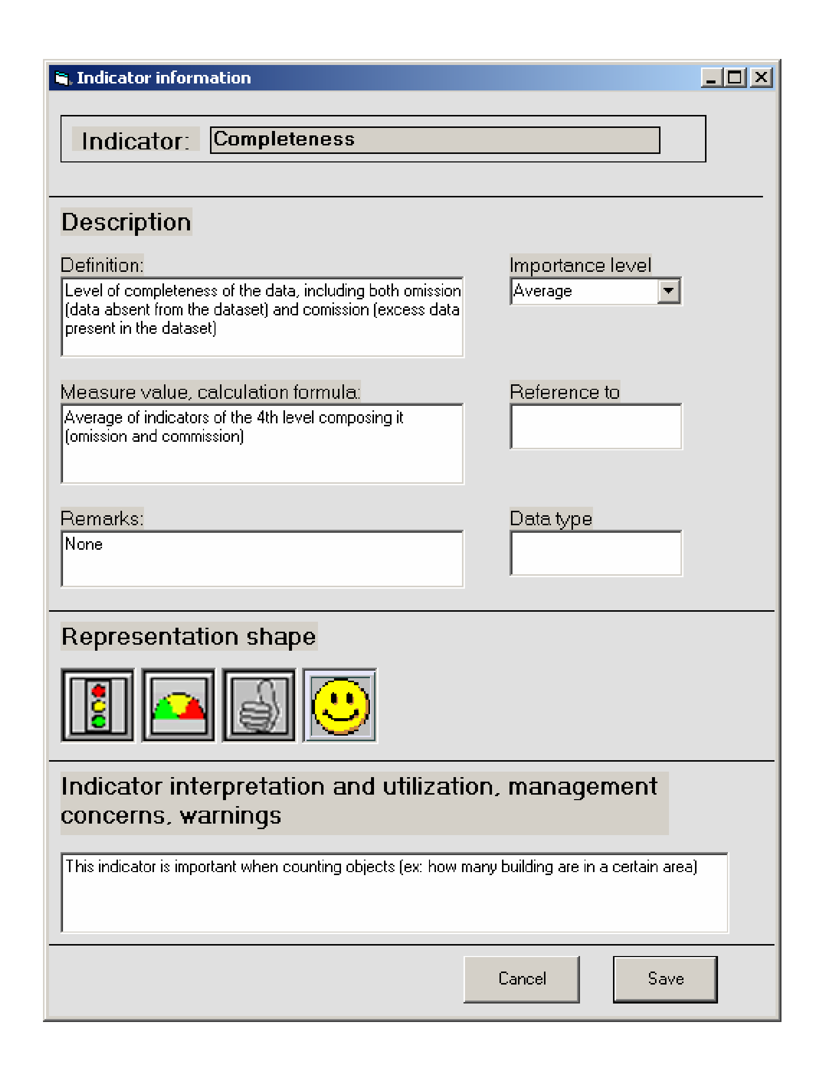


Figure 13 : Exemple de fiche descriptive d'un indicateur de qualité.

3.5.3 Prototype du système MUM

Un prototype du système MUM (Manuel à l'Usager Multidimensionnel) a été développé afin de tester l'approche de communication de la qualité sous la forme d'indicateurs. Le prototype a été programmé en orienté-objet, utilisant des objets de différentes applications, et se base principalement sur trois technologies : SQL Server, GeoMedia et Proclarity. Une base de données multidimensionnelle gérant les informations de qualité a été implantée avec le serveur OLAP SQL Server/ Analysis Services de Microsoft. Le modèle de données utilisé est décrit dans le chapitre 4. Les fonctionnalités cartographiques du prototype (*zoom*

in, *out*, cartes thématiques, etc.) ont été développées avec des objets du logiciel GeoMedia Professional 5 d'Intergraph. Les fonctionnalités OLAP, permettant à l'utilisateur de naviguer dans une base de données multidimensionnelle, ont utilisé des objets du logiciel OLAP-client Proclarity 5. Les données utilisées dans le prototype sont un extrait de la Base Nationale de Données Topographiques du Canada (BNDT) pour le secteur de la ville de Sherbrooke (Québec, Canada). Ces données incluent les routes, bâtiments principaux, rivières, etc. pour des zones de qualité variable.

Pour tester le prototype, les indicateurs proposés par le système sont principalement basés sur la norme internationale ISO 19113 (Principes de qualité) et 19115 (Métadonnées). Les indicateurs de qualité sont gérés de façon hiérarchique selon une dimension dans la base de données multidimensionnelle (cf. chapitre 5). Les indicateurs détaillés sont basés sur une ou plusieurs métadonnées et ceux de plus hauts niveaux sont des agrégations des indicateurs les composant. Pour le prototype, seuls des indicateurs de risque ont été créés, la qualité étant communiquée sous une forme qualitative utilisant différentes représentations telles que des feux de circulation (vert/orange/rouge), *smiley*, etc.

Le prototype offre différentes fonctionnalités telles que :

- La sélection par l'usager d'indicateurs prédéfinis, stockés hiérarchiquement dans une base de données MS-Access. Les indicateurs sélectionnés sont alors affichés dans le tableau de bord;

- La définition d'un profil minimal de l'utilisateur incluant entre autres sa tolérance face au risque et les indicateurs qu'il a sélectionnés;

- La visualisation de fiches descriptives pour chacun des indicateurs présentant leur définition, type de représentation, mode de calcul, etc. (cf. Figure 13);

- La visualisation des indicateurs dans un tableau de bord pouvant inclure jusqu'à 9 indicateurs plus un indicateur global (cf. Figure 14). Ces indicateurs ont été sélectionnés par l'utilisateur parmi une liste hiérarchique d'indicateurs prédéfinis. L'indicateur global représente une agrégation des valeurs des indicateurs sélectionnés, la méthode d'agrégation (ex. maximum, moyenne) dépendant du profil défini par l'utilisateur. L'indicateur global présente une vue générale de la concordance entre la qualité interne des données et les besoins exprimés par les utilisateurs. Utilisant une symbologie de type feu de circulation,

une lumière verte signifie qu'il peut manipuler les données sans risque apparent. Des lumières jaunes ou rouges l'encouragent à explorer les indicateurs le composant;

- La visualisation cartographique des indicateurs, les valeurs de qualité étant associées à chaque entité géométrique (cf. Figure 14). L'utilisateur doit identifier l'indicateur de qualité qu'il désire représenter et chaque objet de la carte prend alors la valeur de qualité qui leur est associée (carte thématique de la qualité utilisant les couleurs vert/jaune/rouge);

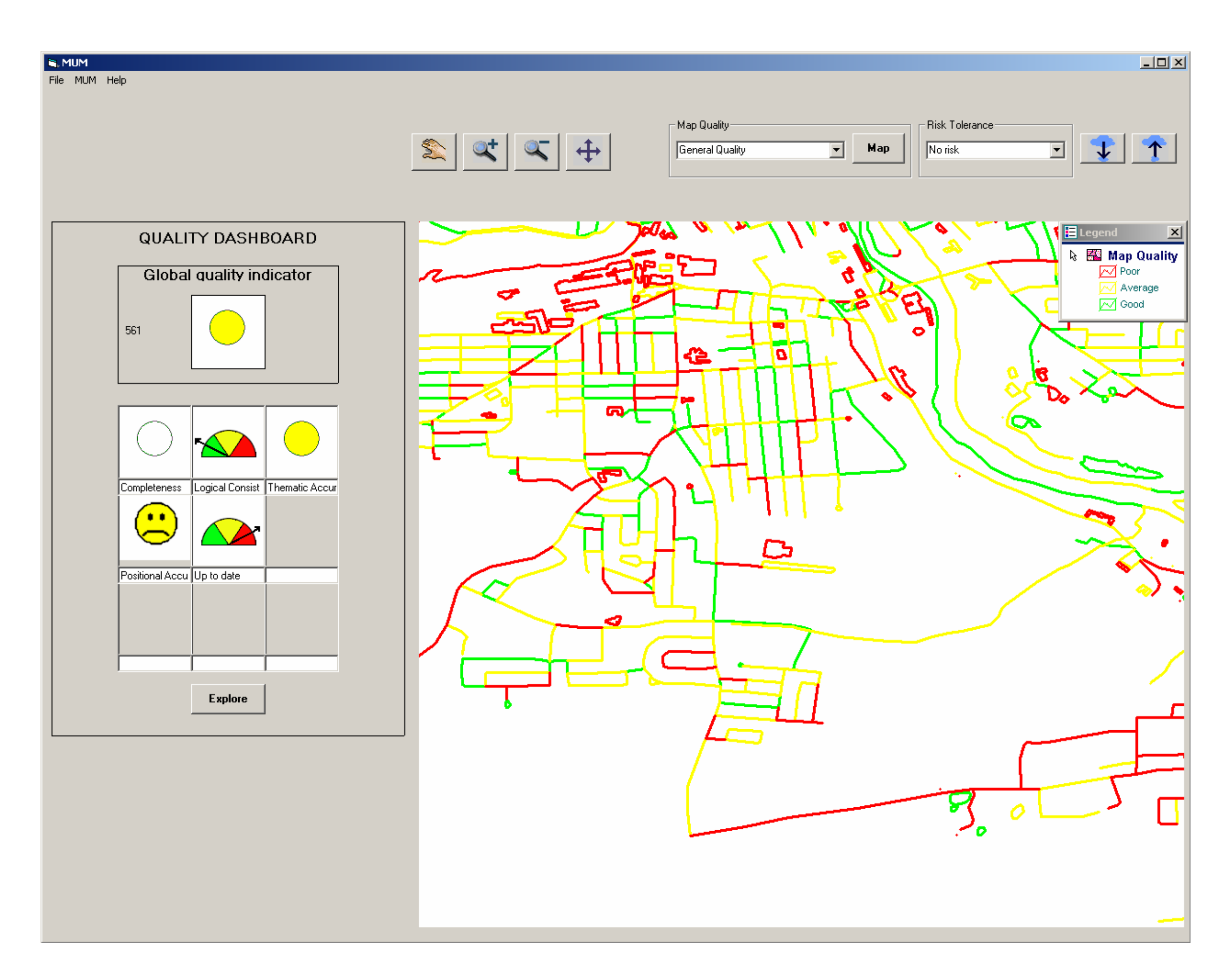


Figure 14 : Interface cartographique du MUM avec tableau de bord et indicateurs (gauche) et représentation cartographique de la qualité (droite). La symbologie vert/jaune/rouge est représentée ici par des niveaux de gris (de gris clair à foncé respectivement).

- La possibilité pour l'utilisateur d'utiliser des fonctions de type OLAP telles que Drill-Down et Roll-Up afin de naviguer dans les données multidimensionnelles à différents

niveaux de détails (ex. visualiser la qualité globale du jeu de données, puis la qualité des routes uniquement, et enfin la qualité d'une seule route). Ces outils permettent également de visualiser les indicateurs de qualité à différents niveaux de détails à l'intérieur de la hiérarchie d'indicateurs (cf. Figure 15). La Figure 15 présente un indicateur et les sous-indicateurs le composant. L'utilisateur peut utiliser les opérateurs de forage OLAP afin de visualiser un niveau plus détaillé ou plus général de la hiérarchie;

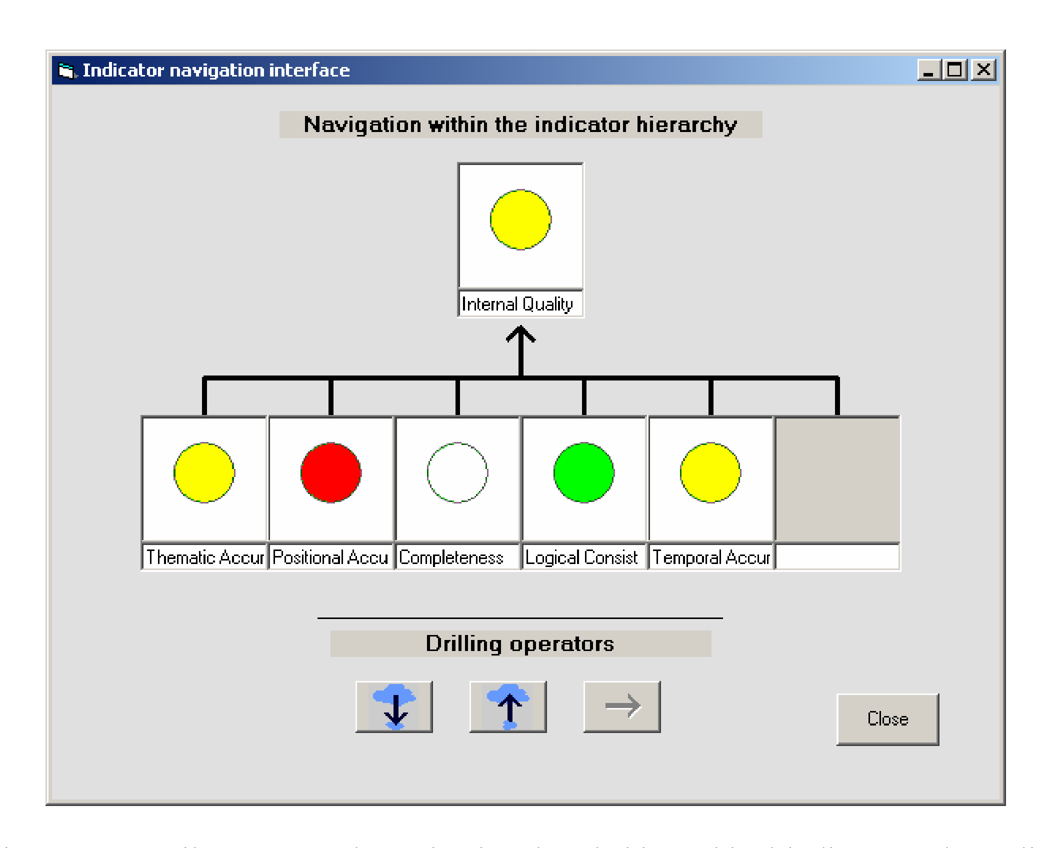


Figure 15 : Outil permettant la navigation dans la hiérarchie d'indicateurs de qualité.

Le tableau de bord permet aux utilisateurs de naviguer dans la hiérarchie des indicateurs de qualité à différents niveaux de détails grâce à des fonctions OLAP. Il est alors possible de visualiser l'information de qualité, de l'indicateur global aux données sources servant au calcul des indicateurs, sans avoir de surcharge d'information.

3.6 Conclusion et perspectives

Cet article présente une nouvelle approche permettant de communiquer l'information relative à la qualité des données géospatiales dans le but de réduire les risques de mauvaises utilisations. Afin de ne pas surcharger les utilisateurs d'informations et de supporter adéquatement leur processus de décision, cette approche préconise l'utilisation de tableaux de bord et d'indicateurs de qualité intégrés dans l'interface du SIG. La qualité étant ici définie comme « l'adéquation à l'utilisation », l'information fournie aux utilisateurs compare les attentes de ceux-ci aux spécifications et caractéristiques internes des données. Cette information relative à la qualité des données peut être basée sur les métadonnées ou toute autre source d'information sur la qualité. L'information sur la qualité est alors communiquée à l'utilisateur sous la forme d'indicateurs de statut ou de risque que celui-ci peut sélectionner, modifier au besoin, puis consulter à différents niveaux de détails. Étant donné l'hétérogénéité spatiale de l'information sur la qualité, des outils permettant une visualisation cartographique de la qualité sont également proposés. Cette approche fournit aux utilisateurs de SIG des outils qui leur permettent d'identifier rapidement des divergences potentielles entre leurs besoins tels qu'exprimés et la qualité des données telle que documentée. Une telle approche peut être intégrée dans des outils SIG ou dans d'autres outils de visualisation cartographique (ex. SOLAP), soit comme un outil de gestion de la qualité à part entière, soit comme une composante de l'outil de visualisation parmi d'autres, pouvant être activée au besoin par l'usager.

Les métadonnées définies par les organismes de normalisation et actuellement fournies par les producteurs de données sont nécessaires pour permettre la création des indicateurs, mais présentent des limites dans leur format actuel. En effet, de nombreuses métadonnées utilisent des textes libres pour décrire les données, ce type de format étant difficilement manipulable automatiquement. De plus, les métadonnées offrent la plupart du temps des descriptions au niveau du jeu de données uniquement. Pour tirer le maximum de bénéfices du système MUM, les données devraient décrire les objets à un niveau de détail plus fin afin de mieux souligner l'hétérogénéité spatiale, temporelle ou descriptive de la qualité. Beaucoup de jeux de données n'ont pas de métadonnées ou ont des métadonnées sommaires. Toutefois, pas d'information est en soi une information utile à l'utilisateur des

données, lui indiquant que les données sont peu documentées et que leur utilisation peut donc être délicate. L'utilisateur peut alors décider de réduire son incertitude en acquérant des informations complémentaires sur les jeux de données ou de travailler avec ces données en absorbant ainsi l'incertitude résiduelle.

Remerciements

Ce travail est financé par le Ministère de la Recherche, de la Science et de la Technologie du Québec dans le cadre de la collaboration avec le projet européen REVIGIS, le Centre de Recherche en Géomatique (CRG) et l'Université Laval. Nous remercions également le Centre d'Information Topographique de Sherbrooke (CIT-S) de Géomatique Canada pour leur support ainsi que des évaluateurs anonymes pour leurs commentaires.

3.7 Bibliographie

Agumya A., Hunter G. J., « Determining fitness for use of geographic information », *ITC Journal*, vol. 2, nº 1, 1997, p. 109-113.

Beard K., « Use error: the neglected error component », *Proceedings of AUTO-CARTO 9*, Baltimore, Maryland, mars 1989, p. 808-817.

Bédard Y., « Uncertainties in Land Information Systems Databases », *Proceedings of Eighth International Symposium on Computer-Assisted Cartography*, Baltimore, Maryland, 29 mars - 3 avril 1987, American Society for Photogrammetry and Remote Sensing et American Congress on Surveying and Mapping, p. 175-184.

Blackmore M., « High or Low Resolution? Conflicts of Accuracy, Cost, Quality and Application in Computer Mapping », *Computers & Geosciences*, vol. 11, nº 2, 1985, p. 345-348.

Buttenfield B. P., « Representing Data Quality », *Cartographica*, vol. 30, nº 2-3, 1993, p. 1-7.

Curry M. R., *Digital places: Living with Geographic Information Technologies*, London & New-York, Routeledge, 1998.

Duckham M., McCreadie J., « An intelligent, distributed, error-aware OOGIS », *Proceedings of 1st International Symposium on Spatial Data Quality*, Hong Kong, 18-20 juillet 1999, p. 496-506.

Duckham M., McCreadie J. E., « Error-aware GIS Development ». *Spatial Data Quality* (W. Shi, P. F. Fisher et M. F. Goodchild, Eds), Taylor & Francis, London, 2002, p. 63-75.

Elshaw Thrall S., Thrall G. I., « Desktop GIS software ». *Geographical Information Systems* (P. A. Longley, M. F. Goodchild, D. J. Maguire et D. W. Rhind, Eds), John Wiley & Sons, New-York, 1999, p. 331-345.

Epstein E. F. Hunter G. J., Agumya A., « Liability insurance and the use of geographical information », *International Journal of Geographical Information Science*, vol. 12, nº 3, 1998, p. 203-214.

Faïz S. O., *Systèmes d'Informations Géographiques: Information Qualité et Data Mining*, Tunis, Éditions C.L.E., 1999.

Fernandez A., *Les nouveaux tableaux de bord des décideurs*, Paris, Éditions d'organisation, 2000.

FGDC, Content Standard for Digital Geospatial Metadata Workbook version 2, 2000.

Fisher G., « User Modeling in Human-Computer Interaction », *User Modeling and User-Adapted Interaction*, vol. 11, 2001, p. 65-86.

Fisher P., « Models of uncertainty in spatial data ». *Geographical Information Systems* (P. A. Longley, M. F. Goodchild, D. J. Maguire et D. W. Rhind, Eds), John Wiley & Sons, Inc., 1999, p. 191-205.

Frank A. U., « Metamodels for Data Quality Description ». *Data Quality in Geographic Information - From Error to Uncertainty* (M. F. Goodchild et R. Jeansoulin, Eds), Editions Hermès, 1998, p. 192.

Gan E., Shi W., « Error Metadata Management System ». *Spatial Data Quality* (W. Shi, P. F. Fisher et M. F. Goodchild, Eds), Taylor Francis, London and New York, 2002, p. 336.

Gervais M., *Pertinence d'un manuel d'instructions au sein d'une stratégie de gestion du risque juridique découlant de la fourniture de données géographiques numériques*, Thèse de doctorat, Université Laval, Québec, 2004.

Goodchild M. F., « Sharing Imperfect Data ». *Sharing Geographic Information* (H. J. Onsrud et G. Rushton, Eds), Rutgers University Press, New Brunswick, NJ, p. 413-425, 1995.

Goodchild M. F., Kemp K. K., NCGIA Core Curriculum in GIS, National Center for Geographic Information and Analysis, University of California, Santa Barbara CA, 1990.

Harvey F., « Quality Needs More Than Standards ». *Data Quality in Geographic Information - From Error to Uncertainty* (M. F. Goodchild et R. Jeansoulin, Eds), Editions Hermès, 1998, p. 192.

Hunter G. J., « Managing uncertainty in GIS ». *Geographical Information Systems* (P. A. Longley, M. F. Goodchild, D. J. Maguire et D. W. Rhind, Eds), John Wiley & Sons, Inc., 1999, p. 633-641.

Hunter G. J., « Spatial Data Quality Revisited ». *Proceedings of GeoInfo 2001 Symposium*, Rio de Janeiro, Brésil, 4-5 octobre 2001.

Hunter G. J., Reinke K. J., « Adapting Spatial Databases to Reduce Information Misuse Through Illogical Operations », *Proceedings of 4th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences (Accuracy 2000)*, Amsterdam, juillet 2000, p. 313-319.

ISO-TC/211, Geographic Information - Quality principles 19113, 2003.

Kaplan R., Norton D., « The balanced scorecard: Measures that Drive Performance », *Harvard Business Review*, vol. 70, nº 1, 1992, p. 71-79.

Klein G., *Sources of Power - How people make decisions*, Cambridge, Massachusetts, MIT Press, 1999.

Krek A., Frank A. U., « Optimization of Quality of Geoinformation Products », *Proceedings of Proceedings of 11th Annual Colloquium of the Spatial Information Research Centre, SIRC'99*, Dunedin, New Zealand, 13-15 décembre, 1999, Department of Information Science, University of Otago, p. 151-159.

Lardon S., Maurel P., Piveteau V., ed., 2001. *Représentations spatiales et développement territorial*, Éditions Hermès.

Longley P. A., Goodchild M. F., Maguire D. J., Rhind D. W., ed., 1999. *Geographical Information Systems*, John Wiley & Sons.

Martinet B., Marti Y.-M., *L'intelligence économique*, Éditions d'Organisation, 2001.

Miller G. A., « The Magical Number Seven, plus or minus two: Some limits on our capacity for processing information », *The Psychological Review*, vol. 63, 1956, p. 81-97.

Miller H. J., Han J., *Geographic Data mining and Knowledge Discovery*, Taylor & Francis, 2001.

Mintzberg H., *The structuring of organisations*, Englewood Cliffs, Prentice-Hall, 1979.

Monmonier M., « A Case Study in the Misuse of GIS: Siting a Low-Level Radioactive Waste Disposal Facility in New-York Sate », *Proceedings of Conference on Law and Information Policy for Spatial Databases*, Tempe (USA), 1994, p. 293-303.

Plan Canada, Sustainable community indicators program, vol 39, nº 5, 1999.

Qiu J., Hunter G. J., « A GIS with the Capacity for Managing Data Quality Information ». *Spatial Data Quality* (W. Shi, M. F. Goodchild et P. F. Fisher, Eds), Taylor & Francis, London, 2002, p. 230-250.

Reinke K. J., Hunter G. J., « A Theory for Communicating Uncertainty in Spatial Databases ». *Spatial Data Quality* (W. Shi, P. F. Fisher et M. F. Goodchild, Eds), Taylor & Francis, London, 2002, p. 77-101.

Rivest S., Bédard Y., Marchand P., « Towards Better Support for Spatial Decision Making: Defining the Characteristics of Spatial On-Line Analytical Processing (SOLAP) », *Geomatica*, vol. 55, nº 4, 2001, p. 539-555.

Roche S., *Les enjeux sociaux des systèmes d'information géographique - le cas de la France et du Québec*, Éditions L'Harmattan, 2001.

Shannon C. E., « A Mathematical Theory of Communication », *The Bell System Technical Journal*, vol. 27, 1948, p. 379-423.

Timpf S., Raubal M., Kuhn, W., « Experiences with Metadata », *Proceedings of Symposium on Spatial Data Handling, SDH'96, Advances in GIS Research II*, Delft, The Netherlands, 12-16 août 1996, IGU, p. 12B.31 - 12B.43.

Voyer P., *Tableaux de bord de gestion et indicateurs de performance*, Presse de l'Université du Québec, 2000.

Chapitre 4 : Gestion de l'information sur la qualité des données

Multidimensional management of geospatial data quality information for its dynamic use within Geographical Information Systems

R. Devillers, Y. Bédard et R. Jeansoulin

Photogrammetric Engineering and Remote Sensing

(sous presse – publication prévue: février 2005)

4.1 Résumé de l'article

Les métadonnées actuellement distribuées devraient permettre aux usagers d'évaluer la qualité (*fitness for use*) des données géospatiales, réduisant ainsi les risques de mauvaise utilisation des données. Toutefois, les métadonnées présentent des limitations et demeurent largement inutilisées. Il existe toujours un besoin de fournir aux utilisateurs des informations sur la qualité de manière plus compréhensible. Cette recherche a pour objectif de communiquer de façon dynamique l'information sur la qualité de façon rapide et intuitive afin de réduire la méta-incertitude qu'ont les utilisateurs concernant la qualité des données

géospatiales et ainsi réduire les risques de mauvaise utilisation des données. Une telle solution nécessite un modèle de données capable de supporter des informations hétérogènes sur la qualité à différents niveaux d'analyse. A l'aide d'une approche basée sur des bases de données multidimensionnelles, cet article propose un cadre conceptuel nommé QIMM (*Quality Information Management Model*) reposant sur des dimensions et des mesures de la qualité. Ce modèle permet à un utilisateur de naviguer facilement et rapidement dans l'information décrivant la qualité grâce à un client SOLAP (*Spatial On-Line Analytical Processing*) associé à une application SIG. Le potentiel du QIMM est illustré par des exemples et un prototype. Par la suite, des manières de communiquer la qualité des données aux utilisateurs sont explorées.

4.2 Abstract

Today metadata should help users to assess the quality (fitness for use) of geospatial data, in order to reduce the risks of data misuse. However, metadata present limitations and remain largely unused. There still exists a need to provide information to users about data quality in a more meaningful way. This research aims to dynamically communicate quality information to the users in a rapid and intuitive way in order to reduce user meta-uncertainty related to geospatial data quality and then reduce the risks of data misuses. Such a solution requires a data model able to support heterogeneous data quality information at different levels of analysis. Using a multidimensional database approach, this paper proposes a conceptual framework named the Quality Information Management Model (QIMM) relying on quality dimensions and measures. This allows a user to easily and rapidly navigate into the quality information using a SOLAP (Spatial On-Line Analytical Processing) client tied to its GIS application. The potential of the QIMM potential is illustrated by different examples and the presentation of a prototype. Finally we present ways to communicate data quality information to users.

4.3 Introduction

The context in which geospatial data is used has changed significantly during the past decade. Users have now easier access to geospatial data and GIS applications, especially through the web. As the use of GIS applications was formerly almost restricted to geospatial experts, it is

now frequent that users with a limited expertise in the geospatial domain use geospatial data. Although this is a positive evolution in general, one problem has emerged: today's typical geospatial data users have less knowledge in the geographical information domain (Agumya and Hunter 1997; Aalders and Morrison 1998; Curry 1998). Consequently, their knowledge about the risks related to the use of geospatial data is limited (Goodchild 1995; Agumya and Hunter 1997; Curry 1998; Elshaw Thrall and Thrall 1999). In that sense, Goodchild (1995) argues that "GIS is its own worst enemy: by inviting people to find new uses for data, it also invites them to be irresponsible in their use". This sometimes leads to faulty decisions based on these data, possibly having significant social, political or economical consequences, several examples being discussed in the literature (Beard 1989; Monmonier 1994; Curry 1998; Agumya and Hunter 2002; Gervais 2004). In order to reduce the risks of misuse, geospatial data producers spend a lot of resources on documenting their datasets to inform the users about the datasets' specifications and quality. Amongst these documents, metadata (*i.e.* data about data) provide information on several aspects of the datasets, such as data producer identification, spatial reference systems, lineage, definition of features or attributes and data quality, to name a few (FGDC 2000; ISO-TC/211 2003). However, metadata are defined in the literature as producer-oriented, offering only limited benefits to the users who want to assess the fitness of the data for their use (Frank 1998; Harvey 1998). In fact, experience shows that metadata do not reach their information goal for non-expert users and are also difficult to understand by many expert users (Timpf *et al.* 1996; Frank 1998; Harvey 1998). Understanding and reaching conclusions, that could be used in Court for example, about the quality of geospatial data rapidly becomes an unmanageable task when one wants to take into consideration the various heterogeneities (spatial, temporal, thematic, acquisition and other) found in a dataset. Consequently, metadata related to data quality usually remain unused by non-expert as well as by experts, even with the best datasets, leaving users in a state of ignorance about the characteristics of the geospatial dataset being used.

As demonstrated by Gervais (2004), an increasing number of geospatial data is intended for general public and must follow legal requirements related to mass-product category. Metadata, as currently provided or defined within international and national standards, do not reach these obligations, especially concerning the requirements of providing easily understandable information as well as information about potential risks of misuse. According

to Gervais, there is a need for a computerized instruction manual that would reduce the risks of misuse by providing to the users of geospatial data information that is easier to understand. Several authors highlighted the need to design such a tool, sometimes identified as "Quality-aware GIS", "Quality GIS" or "Error-aware GIS", that would dynamically take quality information into consideration during data manipulation (visualization, queries, update, etc.) in order for instance to prevent the user from "illogical operations" (Unwin 1995; Hunter and Reinke 2000; Duckham and McCreadie 2002; Qiu and Hunter 2002).

Such systems require to automatically access and use the information related to geospatial data, *i.e.* metadata. Such metadata do not have to be restricted to the metadata identified or provided by different standard organizations or data producers, they can refer to "data about data" in a more general way. However, today's systems have not yet achieved an efficient user-centric management of geospatial data quality information. The goal of this paper is to propose a conceptual framework for the management of geospatial data quality information that aims to go one step ahead of existing solutions.

In the next section, we explain how this research fits into the wider evolution of geospatial data transfer, focusing especially on today's practice of making metadata accessible to users for assessing the fitness for use of their datasets. In Section 4.5, we present the state of the art concerning "what" kind of quality information is available today. We do so by presenting different standards and classifications of data quality information. Section 4.6 presents different hierarchies allowing quality analysis at different levels of detail. Based on the literature, we propose in Section 4.7 a conceptual framework for geospatial quality information management. We describe multidimensional data structures as well as Spatial On-Line Analytical Processing (SOLAP) and discuss their relevance for geospatial quality information management. A framework for a SOLAP model managing data quality information is presented. We then illustrate our approach with different scenarios of user navigation within the quality information model. We finally present our prototype based on the quality information model developed to test the concepts and highlight the impact of such a model on quality information communication.

4.4 Issues about Geospatial data transfer and quality

In the past, geospatial data was typically produced and used within the same organization. Knowledge about data production processes and characteristics, including quality, was more implicit (*i.e.* organizational memory) than explicit (*e.g.* metadata). With the introduction of digital data, the increase of data transfer changed this perspective. The way organizations or people communicate information related to geospatial data evolved in such a way that the transferred information became more accessible or meaningful to a larger group of geospatial data users (cf. Figure 16).

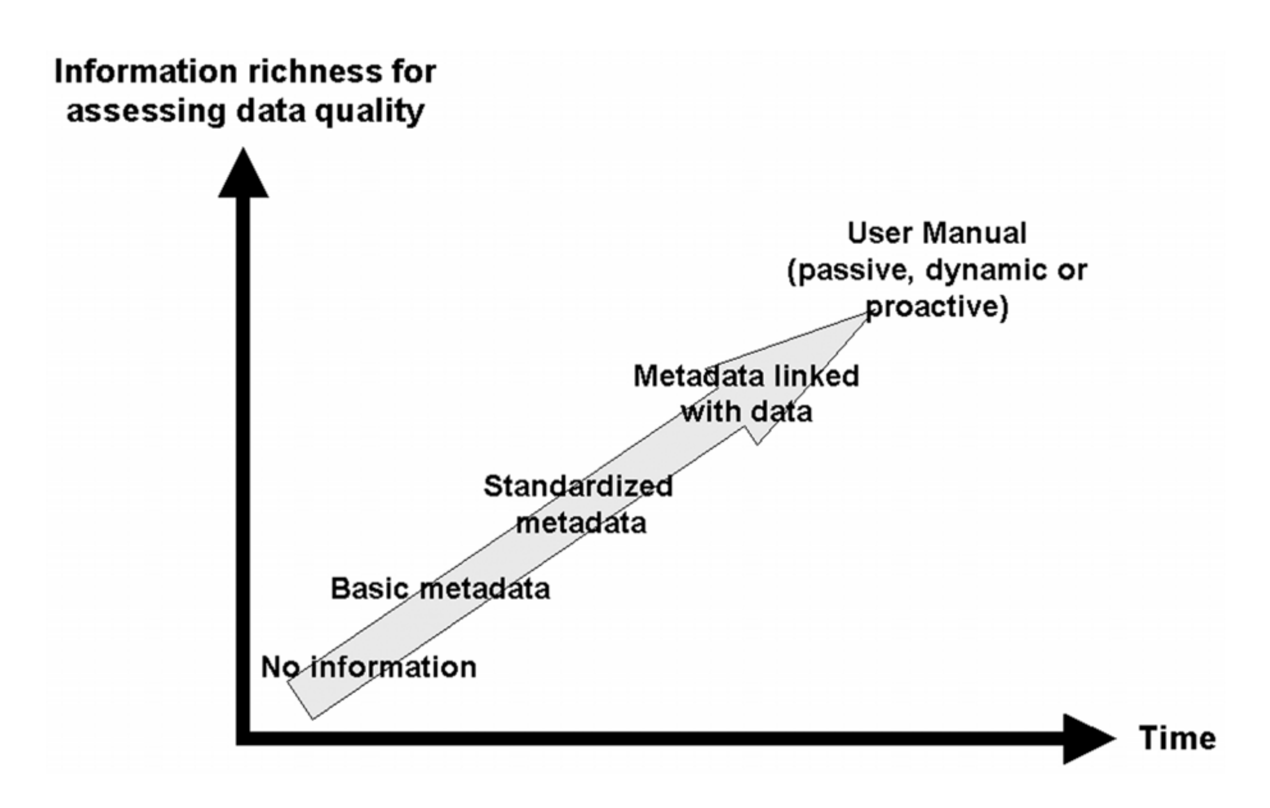


Figure 16 : Evolution of the usefulness of the information communicated to data users for assessing geospatial data quality.

Several stages can be identified:

No quality information: Data is distributed without any associated metadata describing it. This situation is still very frequent and it is not rare to see users specifically asking for the metadata not to be sent, even when they do exist;

Basic quality information: Data producers provide some information when distributing their geospatial datasets, such as dataset reference systems, spatial accuracy or production date. However, this information, not compliant to any standard, is different from one organization to another, describing different characteristics of the datasets at various levels of richness;

Normalized metadata: Local, national or international organizations, such as ISO/TC 211, OpenGIS, FGDC, CGSB/COG or CEN/TC 287, propose geospatial metadata standards in order to homogenize the information shared between the organizations. However, different standards can be used from one organization to another. Often stored in separate text files, these metadata are rarely explicitly associated with their related data, limiting their usefulness for GIS functions (*e.g.* associating uncertainty to distance measurements based on positional accuracy metadata). Furthermore, these standards are more producer-oriented than user-oriented: they are more a formalization of production procedures and tests that are understandable by data acquisition specialists, but they don't provide meaningful information for a general audience useful for decision-making processes;

Metadata linked with data: Metadata provided with datasets are still regularly stored in a text file separate from their data file, without any explicit link between the data and the text file. However, some research works done both in academia and industry are now being performed to strengthen the link between metadata and the data they describe, up to the instance and attribute levels. Beard (1997) mentions that "there is potentially great benefit from an integral association of data with descriptions or measures of its quality. Approaches which separate quality descriptions from the data risk reducing ease of access". Such structured quality information would be accessed more easily by users or software programs but would be more difficult to generalize if the granularity of quality information is very fine. One of the reasons for a tighter link is the need to propagate data updates to metadata. An explicit link between metadata and data would also allow the dynamic use of metadata during data manipulation. Commercial tools such as *ArcGIS ArcCatalog* (ESRI) or *SMMS for Geomedia* (Intergraph) provide a way to manage metadata and dynamically link them to data. However, these tools are still limited in terms of the types of metadata that can be stored and the level of detail of the metadata (*i.e.* metadata are usually stored on the dataset or object class level only).

We suggest a stage further exploiting the metadata structured in the stage 4. This level, exemplified by the MUM (Multidimensional User Manual) project (Devillers *et al.* 2002), provides high-level information or functionalities aiming at reducing the risks of misuse by reducing users' meta-uncertainty when manipulating geospatial data.

The User Manual can be divided into three complementary parts, namely Passive, Dynamic and Proactive User Manual.

Passive User Manual: the passive User Manual is defined as a textual User Manual as usually provided with other goods (*e.g.* medical drugs, electronics), providing different information related to datasets' specifications, possible use and limitations. Such manual can rely on metadata, other information or recommendations provided by data producers, or shared experience from other parties that used these datasets in different contexts. Each manual is contextual, in the sense that it is produced for certain data used in a certain context.

Dynamic User Manual: the dynamic User Manual is designed to be integrated within a GIS interface. Such manual provides users with relevant aggregated information and allows them to navigate at different levels of detail through this information (Devillers *et al.* 2002). Using different levels of detail helps to avoid information overload and to synthesize the quality information. The information provided to the user is either quantitative or qualitative (the latter being more frequent at general levels, while the former more frequent at detailed levels) and would help identify some datasets characteristics that could possibly be risky for the intended use. Doing so requires the comparison of users' expectations and the intrinsic characteristics of geospatial data.

Proactive User Manual: the proactive User Manual is designed to act directly on-the-fly on users' GIS operations in order to avoid some data misuse. This stage requires a database of "illegal operations", as described by Hunter and Reinke (2000). Based on this knowledge and the metadata, the system could also avoid the use of certain functions in some contexts or display a message to warn the user about the possible consequences of the action (*e.g.* restrict data visualization to certain scales based on the data acquisition scale; associate uncertainty to calculations results – *e.g.* distance measurement).

The present paper focuses on Stage 4 presented above, which describes how to link metadata and their associated data to allow the User Manual, or any other "Quality-aware GIS"

functions, to work properly. This work provides the basis on which Stage 5 relies. For the scope of this paper, quality information is defined as any information allowing to assess the quality of a dataset (fitness for use). Hence, quality information includes metadata provided with datasets, but may also include other relevant information or even expert opinions about given data.

4.5 Geospatial Data Quality Characteristics

The definition of a data model allowing the management of geospatial data quality information requires knowing *what* quality information is available and can be integrated into such model. This section provides an overview of the literature related to data quality classifications, looking at both metadata standards and academic research in order to highlight the diversity and similarities of quality classifications, in order to present the limitations of metadata and to justify the QIMM model described in Section 4.7.

Data quality issues have been extensively explored in the geographic information domain for about 20 years. However, there are several definitions of the meaning of "quality". Two trends can be identified in the literature. One restricts quality to datasets' internal characteristics, *i.e.* intrinsic properties resulting from data production methods (*e.g.* data acquisition technologies, data model and storage). This trend is often identified as internal quality. The other trend follows the "fitness for use" definition (Juran *et al.* 1974; Chrisman 1983; Veregin 1999), quality being defined as the level of fitness between data characteristics and users needs. This trend is often identified as external quality. As opposed to the former trend, the latter sees quality as a concept that is relative to the users and usages, neither an independent nor an absolute concept. The assessment of external quality requires information describing the internal quality; the concept of external quality being larger than the internal one. Several classifications of geospatial data quality information have been proposed and can be viewed from two different perspectives: producer and user. The producer point of view generally focuses on internal quality, while the user point of view looks at both internal and external quality.

Several quality characteristics are suggested by standardization organizations and academic researchers for both internal and external qualities. Standardization bodies largely developed the data producer perspective (*e.g.* CEN/TC 287, ICA, ISO/TC 211, OpenGIS, SDTS). They

usually classify data quality into 5 to 7 parameters being: Lineage, Positional accuracy, Attribute accuracy, Semantic accuracy, Temporal accuracy, Logical consistency and Completeness (CEN/TC-287 1994/1995; Guptill and Morrison 1995; FGDC 2000; ISO-TC/211 2003). Each class is usually composed of several sub-classes, but few of these address issues such as accessibility (costs, delays), rights to reproduce (copyright policy), official or legal character of the data, privacy restriction, or any other issues that are needed to assess the fitness for use (from the user's point of view). Table 1 provides an overview of geospatial data quality characteristics identified in standards (*i.e.* CEN, ICA, ISO and SDTS) or by a data producer organization (*i.e.* IGN-France). This table reflects the meaning of quality characteristics (*i.e.* if two organizations have two different names for similar aspects of the quality, they are grouped in the same category).

Table 1 : Examples of data quality characteristics provided by standards or cartographic organizations

	CEN[1]	ICA[2]	IGN[3]	ISO[4]	SDTS[5]
Lineage/Source	X	X		X	X
Spatial/Positional Accuracy	X	X	X	X	X
Attribute Accuracy				X	X
Semantic Accuracy	X	X	X	X	
Completeness	X	X	X	X	X
Logical Consistency	X	X	X	X	X
Temporal Information/Accuracy	X	X		X	

[1] (CEN/TC-287 1994/1995), [2] (Guptill and Morrison 1995), [3] (IGN 1997), [4] (ISO-TC/211 2003), [5] (FGDC 2000)

Table 1 shows that standards and data producers (1) mainly focus on *internal quality* (*e.g.* accuracy, completeness, consistency) aspects and (2) agree, in general, on similar characteristics. Standards are now generally converging to the ISO international standard that may serve as reference for the identification of quality characteristics.

On the other hand, different authors argue that quality assessment defined as “fitness for use” may require information that is not yet included in geospatial metadata standards. They suggest to consider quality characteristics in the wider approach of external quality (*i.e.* quality in the context of use) in addition to internal quality. For instance, Aalders and Morrison (1998) add to the ISO criteria information related to data usage, being previous use

of a dataset by other users for various applications (*i.e.* organization that has used the dataset, type of usage and its perceived fitness, possible constraints or limitations during the use). Bédard and Vallière (1995) bring other characteristics such as legitimacy (legal or *de facto*) and accessibility (costs, delays, easiness to obtain) of the data. Working on data quality issues in general (*i.e.* not restricted to geospatial data), Wang and Strong (1996) identified several characteristics based on a large survey among data users, grouped into four categories: Intrinsic (*e.g.* believability, reputation), Contextual (*e.g.* relevancy, timeliness), Representational (*e.g.* interpretability, ease of understanding) and Accessibility (*e.g.* accessibility, security).

Most of these criteria are not available in today's metadata but would be necessary to help users to assess the fitness for use of datasets for certain applications. For instance, accurate and up-to-date data may not fit for the intended use if the data producer is not recognized (reputation), price is extremely high (cost), time to get them is too long (accessibility) or if data sharing is not permitted (legal issues).

4.6 Geospatial Data Quality Information Hierarchy

The design of a data model allowing the management of geospatial data quality information requires knowing *how* information about data quality is related to the data being described. Quality information can for instance describe a whole dataset quality or only a subset of it (*e.g.* quality of the data related to an object class, quality of the data of a single attribute of an instance). As described by Bédard and Vallière (1995), there are different levels of detail of data quality, also named granularity of data quality. They suggest a method to aggregate quality information from a single data up to the complete dataset. Hunter (2001) identified quality information granularity as one of the main concerns in geospatial data quality research, saying that "data quality suffers generally from being presented at the global level rather than at greatest levels of granularity". Hunter provides several examples illustrating that today's metadata do not provide information at a sufficient level of detail, such as: Positional Accuracy being "Variable", "100m to 1000m" or "+/- 1.5m (urban) to +/- 250m (rural)". The quality of data also varies temporally (*e.g.* +/- 30m before 1992 to +/- 10 meters since 1992 for the more recently covered areas) and thematically (*e.g.* +/- $15000 for residences to +/- 100,000 for stores). These examples illustrate that geospatial data quality

heterogeneity is not adequately recorded in today's metadata to properly assess data quality for the subset of data being used. A description at a more detailed level would allow for quality information to be provided, such as the positional accuracy of a given road, the precision of commercial value of residences in a given area or the level of updateness of building constructions. Although we are well aware that organizations have difficulties complying with today's metadata standards even for the general dataset level, we believe that there exists a need to combine breadth and depth in quality information. The latter can be of varying levels of detail for different features depending on the needs. We also believe, based on Gervais' work (2004), that legal obligations may force data producers and GIS officers to have such detailed information at hand. In fact, this already exists in legally-bounded professional activities such as cadastral surveying, property assessment, road building and other activities where the quality of information is analyzed on a case-by-case basis. Accordingly, this section provides a brief overview of the literature in terms of geospatial metadata levels of detail, looking at metadata standards, academic research and practical illustrations from the Canadian National Topographic Database (NTDB) metadata.

Some authors suggested hierarchies aiming at managing geospatial quality information at different levels of detail (Bédard and Vallière 1995; Faïz 1996; 1999; Qiu and Hunter 1999, 2002)

ISO 19115 standard (2003) provides a framework for encoding metadata for the purpose of search and retrieval, metadata exchange, and presentation. This standard proposes a hierarchy that can be used to store metadata at different levels of detail. This hierarchy may assist in filtering or targeting users' queries to the requested level of detail. The ISO hierarchy goes further than those of Qiu and Hunter's by allowing the association of metadata to attributes (attribute type and instance).

ISO/TC 211 (2003) metadata levels are:

Data series: A series or collection of spatial data, which share similar characteristics of theme, source date, resolution, and methodology. *E.g.* A collection of raster map data captured from a common series of paper maps;

Dataset: Consistent spatial data product instance that can be generated or made available by a geospatial data distributor;

Feature type: Spatial constructs known as features are groups of spatial primitives (0-, 1- and 2 dimensional geometric objects) that have a common identity. *E.g.* All bridges within a dataset;

Feature instance: Spatial constructs (features) that have a direct correspondence with a real world object. *E.g.* The Golden Gate bridge;

Attribute type: Digital parameters that describe a common aspect of grounded spatial primitives (0-, 1- and 2-dimensional geometric objects). *E.g.* Overhead clearance associated with a bridge;

Attribute instance: Digital parameters that describe an aspect of the feature instance. *E.g.* The overhead clearance associated with a specific bridge across a road.

Hierarchies can also be identified within metadata provided by data producers. For instance, the Canadian National Topographic Database (NTDB) metadata has four explicit levels of detail: dataset, metadata polygon, theme and geometric primitive, the latest being directly stored in the data file as attributes.

Therefore, several hierarchies were proposed in the literature. If most of them agree on the general levels (*e.g.* dataset, feature type and feature instance), they often differ at detailed levels. Indeed, some of them do not address the issue of semantic quality (*e.g.* quality of attributes or semantic values), others do not to into account the values of geometric primitives. Regarding the implementation of these hierarchies, some of the approaches are only theoretical while other were tested through prototypes developed using relational databases.

4.7 Multidimensional geospatial data quality management

Juran *et al.* (1974) were the first to define quality as "fitness for use". This definition issued from the quality engineering and management field is now widely recognized in several fields, including the geospatial information community (Chrisman 1983; Veregin 1999). ISO 9000 defines quality as "the totality of characteristics of an entity that bear on its ability to satisfy stated and implied needs". We think that quality is not only the "totality of characteristics of an entity", but rather results from similarity measurements between product specifications and users' needs. In order to highlight this aspect we define quality as the

closeness of the agreement between data characteristics and explicit or implicit needs of a user for a given application. Quality requires taking users' needs into consideration. For this reason, data quality information should not be restricted to the "quality information" section of metadata but should include further information already available in other sections of metadata standards (*e.g.* data coverage or spatial reference systems) or information which is not at all available in today's metadata (*e.g.* accessibility, believability).

4.7.1 Multidimensional Databases – OLAP and SOLAP

In the database field, multidimensional databases such those used in On-Line Analytical Processing (OLAP), are well suited for managing information at different levels of detail. Notice that the term "multidimensional" is used in this paper according to its definition in the database field and is not restricted to spatial and temporal dimensions (x, y, z and t). Multidimensional databases are a component of data warehouses, designed to support data analyses at strategic and tactical levels of organizations. They are opposed to the traditional transactional databases that focus on organization transactions. In the context of data warehouse implementation, multidimensional databases do not replace transactional databases but are complementary by using them as data sources. OLAP systems are tools enabling users to explore, navigate within organizational data structured into a multidimensional database.

OLAP, introduced by Codd (1993), is extensively documented in the database and Business Intelligence fields. CompInfo (2003) defines OLAP tools as "a category of software technology that enables analysts, managers and executives to gain insight into data through fast, consistent, interactive access to a wide variety of possible views of information that has been transformed from raw data to reflect the real dimensionality of the enterprise as understood by the user." OLAP tools support both (1) the management of multidimensional data and (2) the fast retrieval of these data by their users. Their adaptation to the spatial domain, named SOLAP tools by Bédard (1997), can be found in a small number of papers and books (see for example Miller and Han 2001 or Rivest *et al.* 2001) and is emerging today as a powerful complement to GIS (Bédard et al 2003). This is such a SOLAP system that is being used in this project.

SOLAP tools are good candidates to manage geospatial data quality information because:

- of the heterogeneity inherent to geospatial data, which implies that quality information has to be analyzed and managed at different levels of detail;
- of the need to provide contextual aggregated information which is more meaningful to data users. Thus, based on detailed data, SOLAP systems use different ways to aggregate different characteristics, themes, regions, epochs, etc.;

SOLAP tools offer different techniques of data visualization such as matrices, pie charts, histograms, etc. as well as maps;

SOLAP tools are known to be very fast and easy to use. They require no knowledge of query languages. SOLAP delivers rapid "keyboardless navigation" through spatial data and spatial operators at different levels of aggregation (Bédard *et al.* 2003; Marchand *et al.* 2003).

It appears natural to implement our data quality approach into existing decision-support technologies such as SOLAP because of the spatial heterogeneity inherent to geospatial data and of the increased facility to display and explore quality information (cf. maps with tables, statistical charts and semantic trees that can be drilled down or up with a single click of the mouse).

OLAP structures are opposed to the traditional OLTP (On-Line Transactional Processing) structures. The OLTP systems are classical databases implemented to manage transactions (such as bank transactions), and are oriented towards data processing tasks (entering, storing, updating, integrity checking, securing and simple querying of data usually at the level of detail they were collected). In contrast, OLAP systems are oriented towards supporting organizational decision-making by providing aggregated data for both present and historical data (Berson and Smith 1997). OLAP tools rely on multidimensional data models (also called data cubes or hypercubes) which are based on several fundamental concepts such as dimensions, members, measures and facts. "Dimensions" represent the different themes, or thematic axes, from which a user can analyze the data (thus differing from the typical X, Y, Z and T axes commonly used in GIS). Dimensions include members organized into hierarchies. Each dimension can have different levels of detail and each level can include one or several members (*i.e.* nodes in a tree). For instance, a grocery store can use a dimension "Consumer product" including members "Vegetable", "Salad" and "Lettuce" (each member being at a different level of detail). A "measure" is a piece of information (*e.g.* total sales) within a fact

describing the unique combination of members that make this fact. A "fact" is a unique grouping of instantiated measures for the intersection of the different dimensions (*e.g.* the fact "36000$" can be associated to the measure "Total Sales" for the member "Salad" of the dimension "Consumer Product" when intersected with the member "Week 23" of the dimension "Time" and the member "Quebec City" of the dimension "Region"). Different types of models are possible when designing a multidimensional database, such as the star and the snowflake schemas (Berson and Smith 1997). Their implementation can be in typical relational DBMS (called ROLAP), in specialized multidimensional databases (called MOLAP) or in Hybrid multi-tiers architectures (called HOLAP). The selection of the model depends on the type of data and the expected operations.

Different operators (*e.g. drill-down*, *roll-up* and *pivoting*) allow users to navigate into the data. For example, the *Drill-down* operator allows navigating in one dimension from a parent member down to a child member, thus getting more details. *Roll-up* (or *Drill-up*) is the opposite, allowing one to get more global information. These operators do not require any knowledge of database query languages such as SQL, the queries being transparent to the users. They provide instantaneous answers.

Extensions of OLAP to the geospatial data exploration (i.e. SOLAP) have recently been developed in order to support decision-making processes based on geospatial data (Rivest *et al.* 2001; Bédard *et al.* 2003). These systems associate OLAP tools with GIS components to enhance geospatial data visualization and analysis. As geospatial data quality may be highly heterogeneous in space, our research aims at integrating the spatial characteristics of data quality into the QIMM model that could be integrated into traditional GIS or SOLAP tools.

4.7.2 Quality Information Management Model (QIMM)

4.7.2.1 QIMM dimensions

Information about geospatial data quality (*i.e.* quality characteristics) can be organized at different levels of detail along dimensions into an OLAP multidimensional database. We suggest in this paper two dimensions that can structure quality information related to most GIS data (cf. Figure 17).

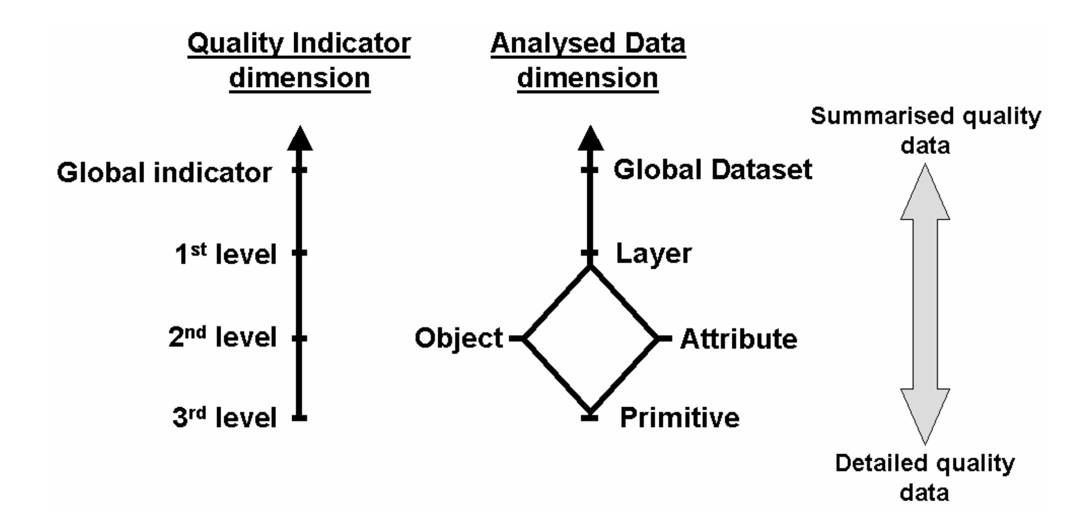


Figure 17 : Quality Information Management Model (QIMM) dimensions and members.

The "Quality Indicator" dimension

Quality indicators provide users with a way to get a quick insight at quality information, and hence contribute to the prevention of potential risks (cf. Chapter 3). Each indicator is based on one or several quality characteristics (cf. Section 4.5) and is implemented as a member of the dimension. In order to avoid information overload, all quality indicators cannot be communicated to data users at the same time. For this reason, they are organized into a hierarchy allowing users to visualize them at different levels of detail. Quality information is aggregated into the dimension hierarchy from the most detailed levels to the more general ones. Members of this dimension (*i.e.* quality indicators) can either provide information regarding the spatial (*e.g.* spatial accuracy), temporal (*e.g.* temporal accuracy) or thematic (*e.g.* attribute accuracy) aspects of the dataset. For instance, members can be horizontal positional accuracy, completeness, date of acquisition or accessibility (see Figure 18 for examples).

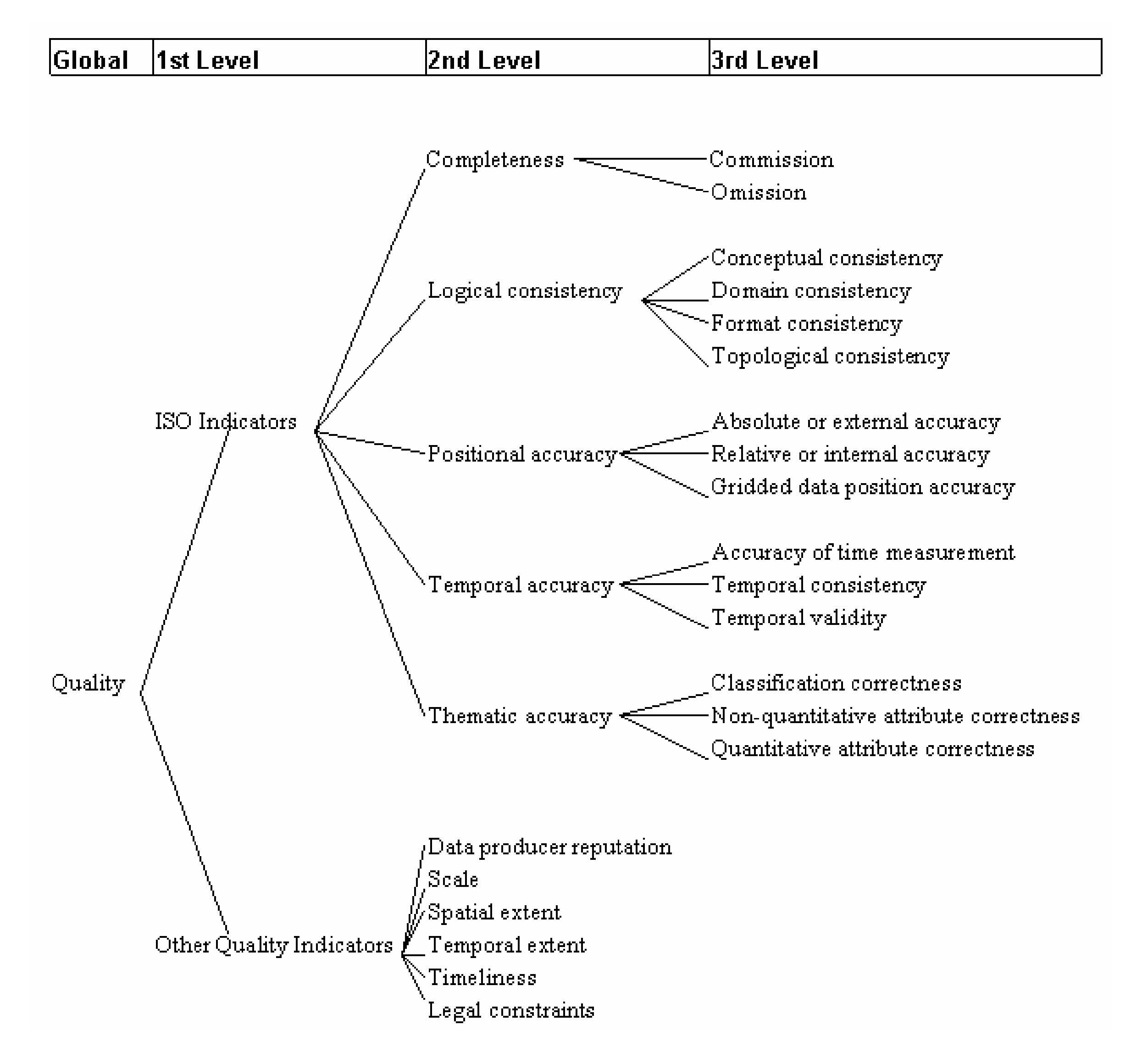


Figure 18 : Example of an indicator hierarchy. Each indicator is a member of the “Quality Indicator” Dimension.

Figure 17 presents four levels of detail as an example but the number of levels of detail can be different according to the user’s preferences. The indicator hierarchy does not have to be balanced. For instance, an indicator located on the second level of detail may not be composed of more detailed indicators on the third and fourth levels. Figure 18 provides an example of an indicator hierarchy mainly based on ISO TC/211 19113 and 19115 standards. Users can define their own indicator hierarchy by selecting pre-defined indicators within a database or defining new ones. The global indicator is the most general quality indicator. It is

an aggregation of all first level indicators and provides an insight on the overall data quality. On the other side, the more detailed level is raw quality information, obtained for instance from metadata.

The "Analyzed Data" dimension

The "Analyzed Data" dimension follows the structure of geospatial data (see an example on Figure 19). In this model, quality information is associated with detailed values (*e.g.* primitive values). Other levels of a dimension hierarchy are either aggregations of the primitive values or raw data if information was only available at more general levels (*e.g.* average quality of lakes without detailed information about the quality of individual lakes). Different aggregation operators available in multidimensional database systems, such as minimum, average or maximum values, can be used, depending on user preferences. Other more complex operators can also be implemented and made available to users (*e.g.* categorizing, above/under, quadratic mean square) to support a more global analysis of quality information. The members of the "Analyzed Data" dimension are grouped in the following levels:

- Primitive – this level can be either geometric (geometric primitives such as points or lines) or semantic (semantic value). For instance, several geometric primitives can compose an object instance, such as a cadastral parcel composed of several lines (each line being defined by at least two points). As these points can be acquired at different dates or using different technologies, the primitives of a same object instance can have different quality levels (*e.g.* quality related to a point located by GPS or to the value "commercial" of the attribute "Type" describing a building);

- Object instance – this level provides all the quality information (geometric and semantic) related to a single instance of object recorded in the dataset (*e.g.* "Beaver Lake" or "Moose Road"). The overall semantic quality for a certain object is an aggregation of the qualities of each data value (*e.g.* aggregated quality of "Road 138");

- Attribute – this level provides the quality related to an object class (or layer) attribute, being an aggregation of primitive value qualities for this attribute (*e.g.* aggregated quality of attribute "house income" for all buildings instances). Notice that only qualities related to the semantics can be associated to the attribute level;

- Layer (or Object Class)– this level provides the aggregation of the quality (geometric and semantic) of all the object instances of a same layer (or class object). A layer can be for instance “Roads”, “Buildings”, “Rivers” or “Parks” (*e.g.* average quality for all lakes);
- Dataset – The dataset includes the quality information (geometric and semantic) related to all the object instances of all data layers. The dataset quality is an aggregation of data layer qualities. A dataset can be for instance a topographic map including lakes, rivers, streets and buildings.

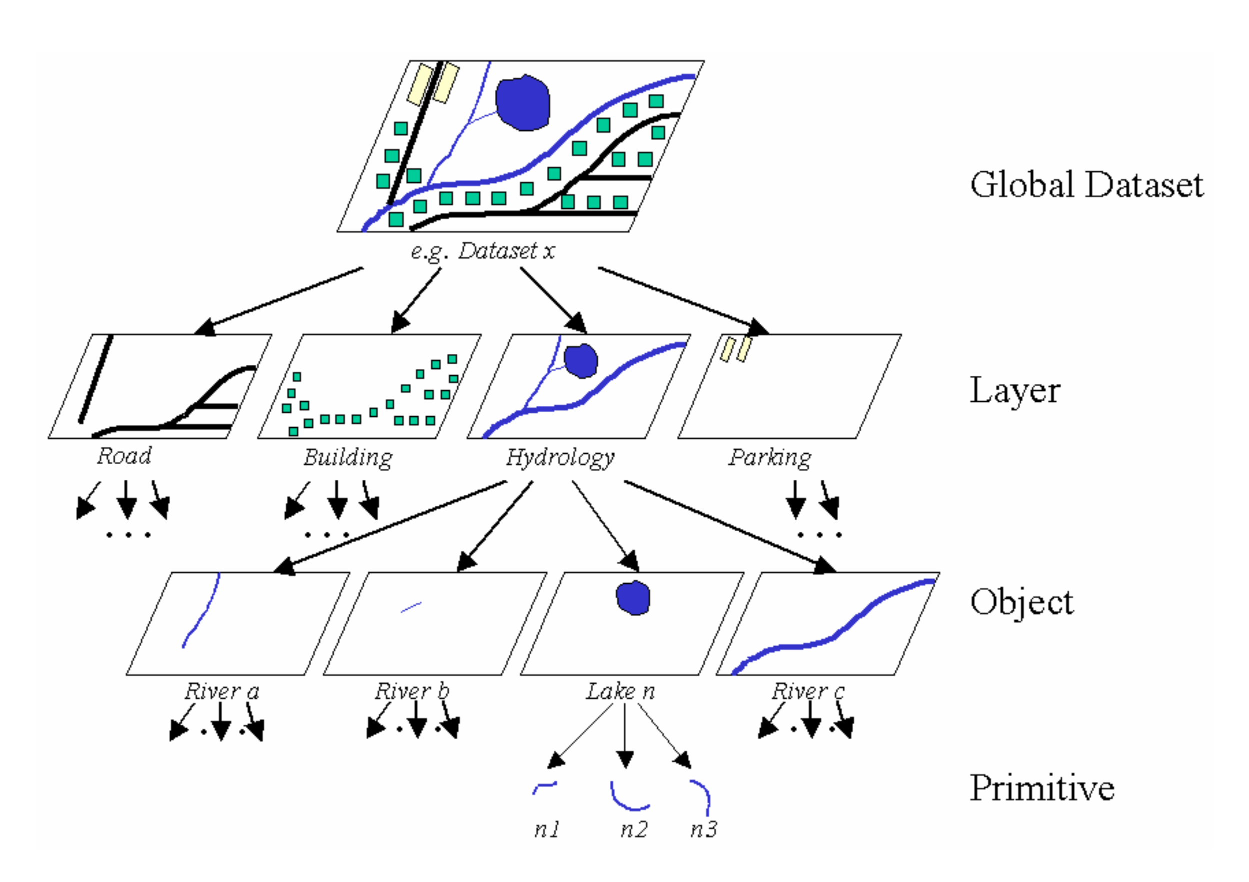


Figure 19 : Example of data hierarchy[17].

The quality of groups of objects can be aggregated from each object’s instance individual qualities. Such a measure can be obtained using spatial queries (*e.g.* what is the overall quality of buildings located in the city “X” or at less than 500 meters from point “Y”?), or queries on semantics (*e.g.* what is the overall quality of buildings of “commercial” type or

[17] This figure was not in the original version of the paper but was added afterwards during the thesis redaction

agricultural parcels of “corn” type). In order to benefit from the SOLAP performance and ease of use, such groups should be predefined.

These levels of the “Analyzed Data” dimension can include one or several members. Members depend on the datasets manipulated by the users (*e.g.* members “Road” and “River” can become members of the level “Layer” when a user adds these data in his GIS environment).

Some intersections between the quality dimensions may be forbidden because of their illogical nature, such as “completeness of a single point” (*e.g.* fire hydrant) or “positional accuracy of the attribute ‘building value’ ”.

4.7.2.2 QIMM measures

Measures are the piece of information describing quality indicators. Measures should describe both internal (spatial or temporal accuracy, completeness, logical consistency, etc.) and external quality characteristics (difference in updateness between a user's expectation and used data, difference in believability, etc.). They can be metadata values or the result of the comparison between metadata values and user’s needs (*e.g.* under, equal or above the needs, represented for instance by green, yellow or red, respectively). As other GIS functions could use quality information stored in the multidimensional database, measures have to be as formalized as possible, avoiding free text for instance, in order to be manipulated more easily by the computer. Quantitative measures are more suitable for data manipulation (*e.g.* aggregation) than qualitative ones. Some measures stored in the multidimensional database can be computed using other measures.

4.7.3 Navigation within the model and quality visualization

Geospatial data users can navigate within the QIMM along both the “Analyzed Data” and the “Quality Indicators” dimensions, moving from a level of detail to another (cf. Figure 20).

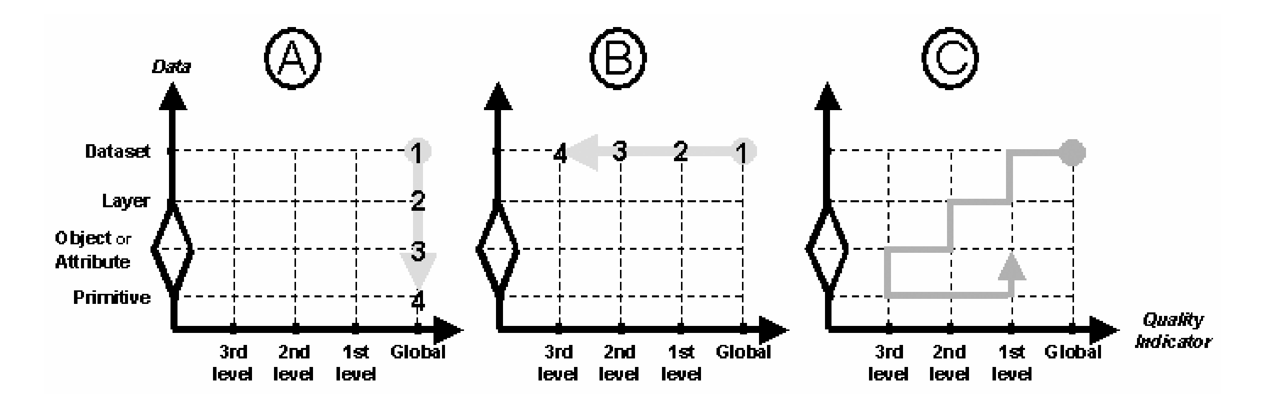
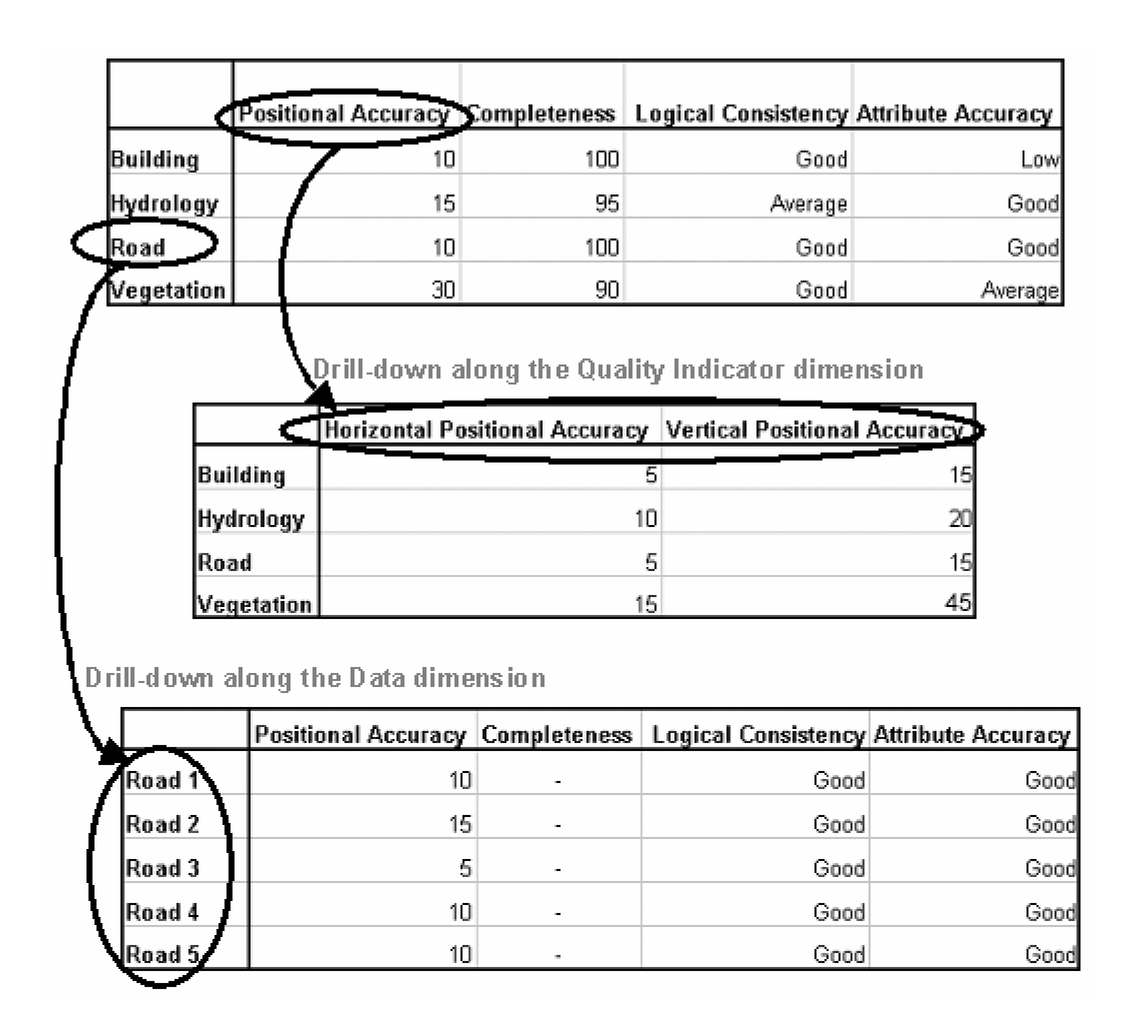


Figure 20 : Examples of user navigation into the quality information along both Quality dimensions

For instance (Figure 20, case A), a user can look at the Global quality indicator for the whole Dataset (position 1: aggregated view of the overall quality for all the objects of the dataset). Then, the user can visualize more details along the "Analyzed Data" dimension using the OLAP drill-down operator, looking at the overall quality for a given layer (*e.g.* position 2: cadastral parcel layer), then for the overall quality of a single object instance (*e.g.* position 3: parcel 147), and finally to the overall quality of parcel 147 geometric data primitive (*e.g.* position 4: one of the corners of the parcel). Another navigation scenario (Figure 20, Case B) explores the quality information along the "Quality Indicator" dimension. A user can then start (position 1) at the Global indicator for the whole dataset, then drill-down to the 1st first level indicator (*e.g.* position 2: spatial quality), visualizing in this case the average quality related to the spatial characteristics of all the objects . The user can then drill-down to the 2nd level indicator (*e.g.* position 3: spatial accuracy) still at the dataset level, and finally to the 3rd level indicator (*e.g.* position 4: horizontal spatial accuracy), being in this case a metadata recommended by ISO and provided into metadata by data providers. Case C (Figure 20) provides an example of a more complex navigation, using successive drill-down and roll-up operations along both dimensions. Such navigation allows a user to follow his line of thought when exploring quality information provided by a fast and easy user interface such as a SOLAP interface.

Figure 21 provides an example of navigation within quality information displayed in a tabular view using drill-down operations along the two quality dimensions. The first drill-down is

performed on the “Quality Indicator” dimension, allowing the user to move from one level of detail to a more detailed level on this dimension. The second one (*i.e.* drill-down on Roads) is performed on the “Analyzed Data” dimension, allowing the user to move from the “Layer” member down to the “Object” member.



	Positional Accuracy	Completeness	Logical Consistency	Attribute Accuracy
Building	10	100	Good	Low
Hydrology	15	95	Average	Good
Road	10	100	Good	Good
Vegetation	30	90	Good	Average

	Horizontal Positional Accuracy	Vertical Positional Accuracy
Building	5	15
Hydrology	10	20
Road	5	15
Vegetation	15	45

	Positional Accuracy	Completeness	Logical Consistency	Attribute Accuracy
Road 1	10	-	Good	Good
Road 2	15	-	Good	Good
Road 3	5	-	Good	Good
Road 4	10	-	Good	Good
Road 5	10	-	Good	Good

Figure 21 : Examples of user navigation in a tabular view using the drill-down operator on the two QIMM dimensions.

Based on the QIMM data structure, users can access different displays of quality information, facilitating their analysis. For instance, indicator values can be displayed in a dashboard, on a map or directly in the descriptive data table (cf. Figure 22). These are examples of possible quality visualization techniques but a wide range of other techniques can benefit from the quality information stored in the QIMM.

- *Dashboard visualization*: Quality indicator values can be displayed in a dashboard (cf. chapter 3), such as dashboards used by many decision-support systems. Indicators can have different representations (*e.g.* number, street light, speed meter, smiley) depending on the type of data to be represented and the user's preference. Figure 22 presents a dashboard including five quality indicators selected by the user because they are relevant in his context. Each indicator value is displayed using the representation selected by the user. The dashboard is displayed into the GIS interface and can be visible or not. These indicators represent quantitative or qualitative values resulting from the comparison of the data characteristics and the user's needs. A User can visualize indicators at different levels of details and can navigate in the indicator hierarchy using OLAP operators (*e.g. drill-down* and *roll-up*).

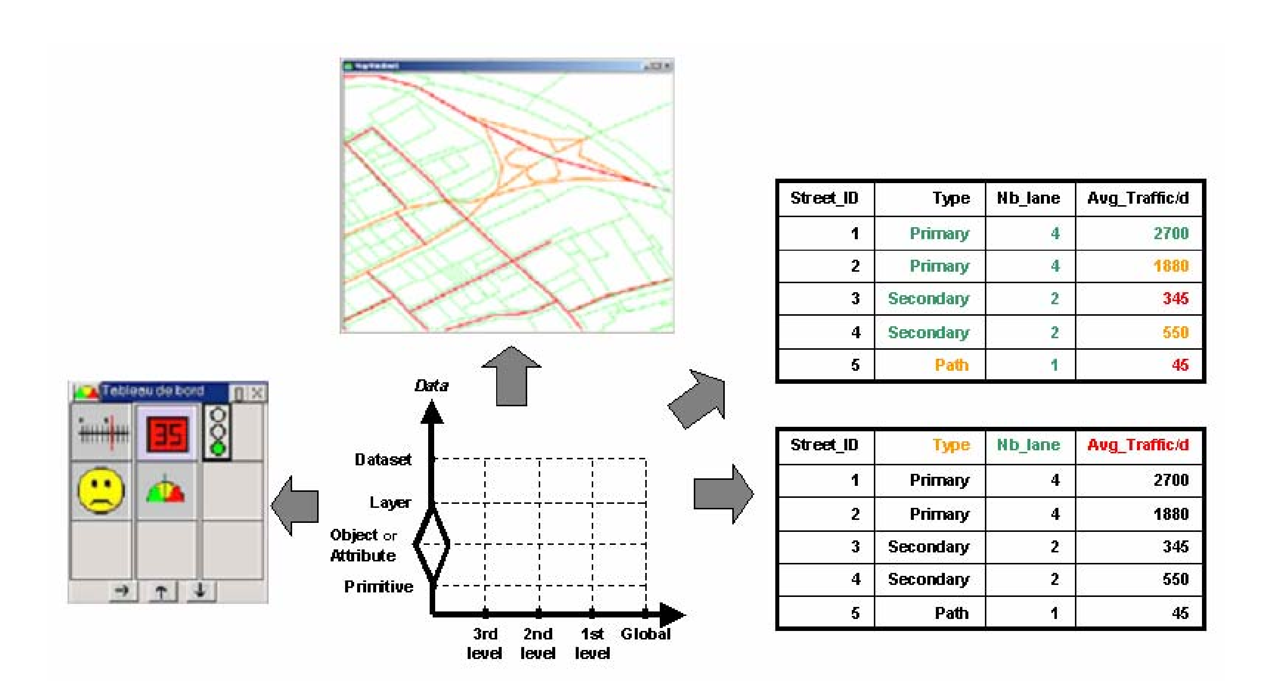


Figure 22 : Possible visualizations of Quality information using the QIMM. Quality information can be for instance displayed in a dashboard (left), on a cartographic base (top), in attribute tables on the individual value level (top right) or on the attribute level (bottom right).

- *Cartographic visualization*: indicator values can be displayed on a cartographic base using different representations (*e.g.* colour, shape, texture). SOLAP operators can allow

the navigation between the levels of detail in a cartographic view (*e.g.* moving from the visualization of a quality indicator for a single road to the visualization of the quality of each road segment of this road). This visualization mode is particularly interesting to get an idea of the spatial heterogeneity of quality information, users being able to rapidly identify the areas of a map having lower quality and the areas having higher quality. Users can also choose the quality parameter they want to visualize (*e.g.* positional accuracy of objects, temporal accuracy).

- *Descriptive data table visualization*: Indicators related to semantic quality, such as attribute accuracy or completeness can be visualized within the data table at different levels of detail. In this way, a user can have a quick insight on the quality of descriptive data contained in a traditional data table as provided by most GIS software. Figure 22 shows the visualization of values for individual data qualities in the first table (for one instance) and an aggregation of values for data qualities at the attribute level in the second table (*i.e.* for all instances).

The visualization techniques used in a SOLAP (*i.e.* maps, tables, statistical charts, semantic tree) allow users to navigate into quality information from one level of detail to another along both “Quality Indicators” and “Analyzed Data” dimensions as shown in the next section.

4.7.4 The MUM prototype

A prototype was developed to test the QIMM model introduced in this paper with a user interface made of a simple dashboard and cartographic visualization. The prototype is based on three main technologies integrated into a single cartographic interface: (1) a multidimensional database storing quality information at different levels of detail into a MOLAP hypercube implemented using Microsoft’s SQL Server/Analysis services, (2) cartographic functionalities using GeoMedia Professional GIS from Intergraph, and (3) OLAP tools enabling a user to navigate into quality information along the two dimensions of the QIMM model, both in tabular and cartographic views, using Proclarity’s OLAP software. The resulting SOLAP prototype was tested with data from the Canadian National Topographic Database (NTDB).

This prototype supports different functionalities such as:

- Managing quality information into a multidimensional database structure using a subset of the QIMM model (from the data level to the object instance level). The QIMM measures are mostly based on quality elements and sub-elements described in the ISO 19113 standard. The QIMM dimensions (*i.e.* data and indicator) were implemented under SQL Server;
- Loading and viewing geospatial data (*e.g.* zoom in, zoom out, pan, fit all). Spatial objects are linked to the quality information stored in the QIMM using a foreign key;
- Visualizing quality information using indicators displayed in a dashboard and on a cartographic display. Indicators are selected by users within an indicator dataset stored in an Access relational database.
- OLAP functions (*e.g. drill-down*, and *roll-up*) allowing users to navigate into quality information along both “Analyzed Data” and “Quality Indicators” dimensions.

Quality information obtained from metadata is transformed into risk levels, based on user-defined tolerance levels. Then, quantitative quality information (*e.g.* 15 meters for positional accuracy) is compared to a user tolerance level (*e.g.* 1 meter) and then transformed into quantitative values for detailed information or into qualitative values such as green/yellow/red streetlight display for lower detailed information. The qualification of quality information uses user-defined thresholds. Other more complex techniques could be used as mentioned in section 4.7.2.1.

Figure 23 shows the main interface of the MUM prototype. This interface is composed of a cartographic view displaying the NTDB dataset, a quality indicator dashboard (located on the left part of the display) and different tools offered to the user (located on the top of the cartographic view). They are from the left to the right: cartographic tools (*e.g. pan*, *zoom in*, *zoom out*, *fit all*), MUM tools (*i.e.* selection of the quality element to be mapped, definition of the user’s tolerance to risk) and some OLAP tools (*i.e. drill-down* and *roll-up*). This example shows the values for six quality indicators selected by the user (commission, omission, up to date, etc.) and for a global quality indicator. General quality (aggregation of all quality indicators) was mapped by the user in order to visualize the spatial heterogeneity of quality at the general level.

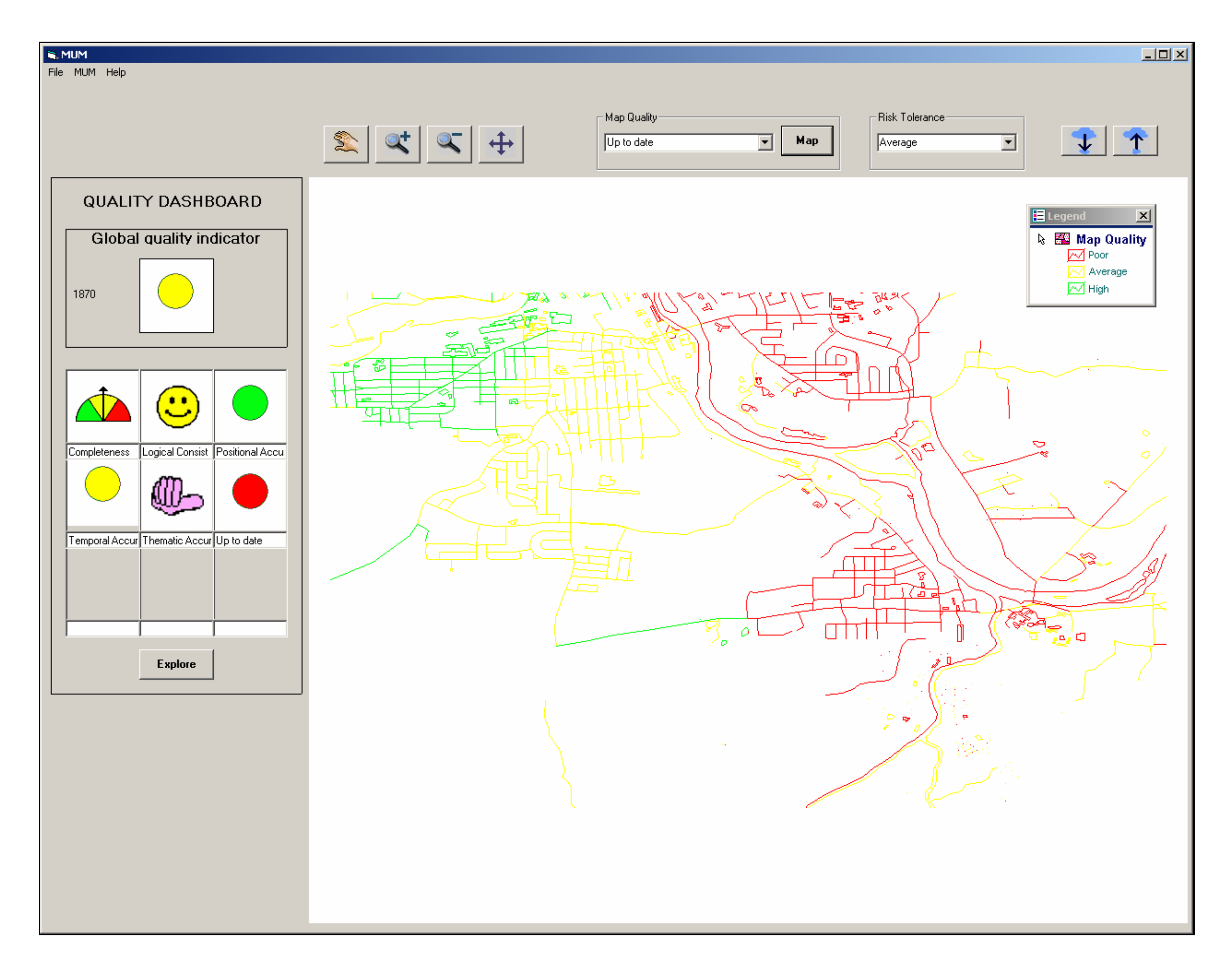


Figure 23 : Prototype using the QIMM model to manage and communicate data quality information

As seen on Figure 23, an important outcome of this approach is to support the spatial variability of quality information. Indeed, because of the heterogeneity of acquisition methods used to acquire geospatial data (*e.g.* Total station, GPS, aerial images), to update them (spatial extent and frequencies, differences in methods), the different objects and geometric primitives contained in a geospatial database can have varying levels of quality. The high level of granularity potentially used for quality information in the QIMM model (down to the geometric or semantic primitives level) allows a very powerful analysis of quality when desired. That is the calculation of quality exclusively for the spatial extent defined or visualized by the users. Hence, quality information displayed to the user is an aggregation of qualities of every object instances located in the user-defined area or in his cartographic view. Different possible aggregation techniques were mentioned earlier in the

paper. Using them, users can get different quality information (*e.g.* spatial accuracy, logical consistency, temporal accuracy) for an area of interest and identify areas having higher quality than others. This allows users to get better information on the spatial heterogeneity of quality information.

4.8 Conclusion and perspectives

This paper provided an innovative approach to manage geospatial data quality information based on a multidimensional data management approach. We first highlighted the need to structure quality information in order to provide meaningful and contextual information to geospatial data users. The concepts of Passive, Dynamic and Proactive Multidimensional User Manuals (MUM) were introduced. We presented different works published by standardization and academic bodies classifying data quality into several categories. Several works that aimed at recording data quality at different levels of detail were afterwards discussed. Based on the literature, we presented a conceptual framework named QIMM, allowing the management of quality information at different levels of detail and using a multidimensional database approach. QIMM dimensions (*i.e.* quality indicators and data) and measures were defined and illustrated. Examples of user navigation into quality information were provided to illustrate this approach. Different kinds of quality information visualization were presented and discussed. Finally, a prototype based on the QIMM model has been presented to test the model and highlight the benefits of such an approach to allow diverse ways to communicate quality information.

This work provides a theoretical framework to manage and communicate to users the heterogeneous quality information at different levels of detail. If it is rather frequent to find papers mentioning that quality is multidimensional, this work is the first attempt to structure quality information using a multidimensional approach and SOLAP tools. The QIMM provides answers to a main issue of the spatial data quality field: the need to manage various quality information at different levels of detail. The model was implemented using a commercial multidimensional database, an OLAP software and a commercial GIS. Such a tool can support users in assessing if the quality of geospatial data is good enough for their needs. In situations where quality information is very heterogeneous and the overall quality assessment too complex for non-expert users, such a tool can help geomatics engineers to

support non-expert users to assess if the quality is sufficient according to their requirements. The QIMM implementation is not restricted to multidimensional databases: it is also useful for spatial data quality management in general using traditional relational databases. The quality information being structured at different levels of detail, it can be exploited by different "Quality-aware GIS" programs (*e.g.* uncertainty management, uncertainty/quality communication and visualization, error buttons). Furthermore, detailed quality information allows the cartographic visualization of the spatial heterogeneity of quality. Finally, providing aggregated information to users helps reducing the risks of misuse by reducing the uncertainty related to data quality. This meta-uncertainty is reduced by both the communication of internal quality information and the communication of risk indicators based on external quality, *i.e.* the difference between internal quality values and users requirements.

Acknowledgements

This work is part of the MUM project (Multidimensional User Manual) and is funded in part by the Canadian Network of Centres of Excellence GEOIDE, the IST/FET program of the European Community (through the REV!GIS project), the Ministère de la Recherche, de la Science et de la Technologie du Québec, the Centre for Research in Geomatics (CRG) and Université Laval. Special thanks to Dr. Jean Brodeur and anonymous reviewers for the critical review of the manuscript and Geomatics Canada CTI-S for their support.

4.9 References

Aalders, H.J.G.L., and J. Morrison, 1998. Spatial Data Quality for GIS, Geographic Information Research: Trans-Atlantic Perspectives, Taylor & Francis, London/Bristol, pp. 463-475.

Agumya, A., and G.J. Hunter, 1997. Determining fitness for use of geographic information, ITC Journal, 2(1):109-113.

Agumya, A., and G.J. Hunter, 2002. Responding to the consequences of uncertainty in geographical data, International Journal of Geographical Information Science, 16(5):405-417.

Beard, K., 1989. Use error: the neglected error component, Proceedings of AUTO-CARTO 9, March, 1989, Baltimore, Maryland, pp. 808-817.

Beard, K., 1997. Representations of Data Quality, Geographic Information Research: Bridging the Atlantic (M. Craglia, and H. Couclelis, editors), Taylor and Francis, pp. 280-294.

Bédard Y, 1997. Spatial OLAP. 2nd Annual R&D Forum, Geomatics IV. Canadian Institute of Geomatics. Montreal, November 13-14th.

Bédard, Y., P. Gosselin, S. Rivest, M.-J. Proulx, M. Nadeau, G. Lebel, and M.-F. Gagnon, 2003. Integrating GIS Components with Knowledge Discovery Technology for Environmental Health Decision Support, International Journal of Medical Informatics, 70(1):79-94.

Bédard, Y., and D. Vallière, 1995. Qualité des données à référence spatiale dans un contexte gouvernemental, Technical report for the Ministère des Ressources Naturelles, Université Laval, Québec, Canada.

Berson, A., and S.J. Smith, 1997. Data Warehousing, Data Mining and OLAP (Data Warehousing / Data Management), McGraw-Hill, New-York, 612 p.

CEN/TC-287, 1994/1995. WG 2, Data description: Quality. Working paper N. 15, August 1994. PT05, Draft Quality Model for Geographic Information, Working paper D3, January 1995.

Chrisman, N.R., 1983. The Role of Quality Information in the Long Term Functioning of a Geographical Information System, Proceedings of International Symposium on Automated Cartography (Auto Carto 6), Ottawa, Canada. pp. 303-321.

Codd, E.F., 1993. Providing OLAP (On-line Analytical Processing) to User-Analysts: An IT Mandate, E. F. Codd and Associates.

Compinfo, 2003. http://www.compinfo-center.com/entsys/olap.htm

Curry, M.R., 1998. Digital places: Living with Geographic Information Technologies, Routeledge, London & New-York, 191 p.

Devillers, R., M. Gervais, Y. Bédard, and R. Jeansoulin, 2002. Spatial Data Quality: From Metadata to Quality Indicators and Contextual End-user Manual, Proceedings of OEEPE-ISPRS Joint Workshop on Spatial Data Quality, March 20-21st 2002, Istanbul.

Duckham, M., and J.E. McCreadie, 2002. Error-aware GIS Development. Spatial Data Quality (W. Shi, P. F. Fisher, and M. F. Goodchild, editors), Taylor & Francis, London, pp. 63-75.

Elshaw Thrall, S., and G.I. Thrall, 1999. Desktop GIS software. Geographical Information Systems (P. A. Longley, M. F. Goodchild, D. J. Maguire, and D. W. Rhind, editors), John Wiley & Sons, New-York, pp. 331-345.

Faïz, S.O., 1996. Modélisation, exploitation et visualisation de l'information qualité dans les bases de données géographique, Ph.D. thesis, Université Paris-Sud.

Faïz, S.O., 1999. Systèmes d'Informations Géographiques: Information Qualité et Data Mining, Tunis, 362 p.

FGDC, 2000. Content Standard for Digital Geospatial Metadata Workbookversion 2.

Frank, A., 1998. Metamodels for Data Quality Description, Data Quality in Geographic Information - From Error to Uncertainty (M. F. Goodchild, and R. Jeansoulin, editors), Editions Hermes, pp. 192.

Gervais, M., 2004. La pertinence d'un manuel d'instruction au sein d'une stratégie de gestion de risque juridique découlant de la fourniture de données géographiques numériques, Ph.D. thesis, Université Laval, Québec.

Gervais, M., R. Devillers, Y. Bédard, and R. Jeansoulin, 2001. GI Quality and decision making: toward a contextual user manual, Proceedings of GeoInformation Fusion and Revision Workshop, April 9-12, Quebec city, Canada.

Goodchild, M.F., 1995. Sharing Imperfect Data. Sharing Geographic Information (H. J. Onsrud, and G. Rushton, editors), Rutgers University Press, New Brunswick, NJ, pp. 413-425.

Guptill, S.C., and J.L. Morrison, 1995. Elements of spatial data quality, Elsevier Science, New York, 202 p.

Harvey, F., 1998. Quality Needs More Than Standards. Data Quality in Geographic Information - From Error to Uncertainty (M. F. Goodchild, and R. Jeansoulin, editors), Editions Hermes, pp. 192.

Hunter, G.J., 2001. Spatial Data Quality Revisited, Proceedings of GeoInfo 2001, 4-5th October, Rio de Janeiro, Brazil, pp.1-7.

Hunter, G.J., and K.J. Reinke, 2000. Adapting Spatial Databases to Reduce Information Misuse Through Illogical Operations, Proceedings of 4th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences (Accuracy 2000), July 2000, Amsterdam, pp. 313-319.

IGN, 1997. Bulletin d'information de l'IGN - Qualité d'une base de données géographique: concepts et terminologie, N. 67.

ISO-TC/211, 2003. Geographic Information - Metadata19115.

Juran, J.M., F.M.J. Gryna, and R.S. Bingham, 1974. Quality Control Handbook, McGraw-Hill, New-York.

Marchand, P., A. Brisebois, Y. Bédard, and G. Edwards, 2003. Implementation and evaluation of a hypercube-based method for spatio-temporal exploration and analysis, Journal of the International Society of Photogrammetry and Remote Sensing (theme issue "Advanced techniques for analysis of geo-spatial data"):accepted for publication.

Miller, H.J., and J. Han, 2001. Geographic Data mining and Knowledge Discovery, Taylor & Francis, 338 p.

Monmonier, M., 1994. A Case Study in the Misuse of GIS: Siting a Low-Level Radioactive Waste Disposal Facility in New-York State, Proceedings of Conference on Law and Information Policy for Spatial Databases, Tempe (AZ) USA, pp. 293-303.

Qiu, J., and G.J. Hunter, 1999. Managing Data Quality Information, Proceedings of International Symposium on Spatial Data Quality, 18-20 July 1999, Hong Kong, pp. 384-395.

Qiu, J., and G.J. Hunter, 2002. A GIS with the Capacity for Managing Data Quality Information. Spatial Data Quality (W. Shi, M. F. Goodchild, and P. F. Fisher, editors), Taylor & Francis, London, pp. 230-250.

Rivest, S., Y. Bédard, and P. Marchand, 2001. Towards Better Support for Spatial Decision Making: Defining the Characteristics of Spatial On-Line Analytical Processing (SOLAP), Geomatica, 55(4):539-555.

Timpf, S., M. Raubal, and W. Kuhn, 1996. Experiences with Metadata, Proceedings of Symposium on Spatial Data Handling, SDH'96, Advances in GIS Research II, August 12-16, 1996, Delft, The Netherlands, pp. 12B.31 - 12B.43.

Unwin, D., 1995. Geographical information systems and the problem of error and uncertainty, Progress in Human Geography, 19:548-549.

Veregin, H., 1999. Data quality parameters, Geographical Information Systems (P. A. Longley, M. F. Goodchild, D. J. Maguire, and D. W. Rhind, editors), John Wiley & Sons, Inc., pp. 177-189.

Wang, R.Y., and D.M. Strong, 1996. Beyond Accuracy: What Data Quality Means to Data Consumers, Journal of Management Information Systems, 12(4):5-34.

Chapitre 5 : Prototype MUM

Quality information system to support experts in their assessment of the fitness for use of geospatial data

R. Devillers, Y. Bédard, R. Jeansoulin, B. Moulin

Soumis le 07/09/2004 au journal[18]

International Journal of Geographical Information Science (IJGIS)

5.1 Résumé de l'article

Les utilisateurs de données géospatiales sont de plus en plus confrontés au problème complexe de l'évaluation de l'adéquation de données à un usage défini. Étant donné la disponibilité croissante de sources de données, les jeux de données sont plus que jamais hétérogènes et complexes à interpréter. L'information décrivant la qualité des données est disponible mais demeure souvent elle-même hétérogène sémantiquement et spatiallement, inaccessible, hermétique, et finit en pratique par être négligée par la plupart des utilisateurs.

[18] Dans l'attente d'une réponse de la revue au moment du dépôt de la thèse

Une personne doit en fait pouvoir développer une expertise solide pour comprendre correctement les métadonnées et évaluer l'adéquation de jeux de données, ou d'extraits de ces jeux, pour des usages spécifiques dans des endroits précis et pour des périodes variables. Une telle tâche complexe peut impliquer des milliers de métadonnées partiellement corrélées. En conséquence, des experts en qualité des données doivent pouvoir s'aider d'outils allant les aider à identifier des problèmes potentiels ainsi que les aider à synthétiser les informations nécessaires pour écrire leur opinion dans un rapport impliquant leur responsabilité professionnelle. Afin de supporter de tels experts dans l'évaluation de l'adéquation à l'utilisation (*fitness for use*), cet article présente une approche visant à mieux gérer et communiquer l'information sur la qualité des données grâce à un ensemble de concepts relié aux bases de données décisionnelles et aux techniques de visualisation. Cette approche repose techniquement sur une combinaison des fonctions d'un SIG avec des technologies d'intelligence décisionnelle (principalement le *On-Line Analytical Processing* - OLAP), afin d'adapter l'approche de tableau de bord exécutif pour fournir des indicateurs interactifs et contextuels décrivant la qualité des données géospatiales. Un prototype nommé MUM (Manuel à l'Usager Multidimensionnel) est présenté afin d'illustrer cette approche.

5.2 Abstract

Geospatial data users increasingly face the complex problem of assessing the fitness of datasets for an intended use. Due to the increasing availability of data sources, datasets are more than ever heterogeneous and complex to interpret. Information describing data quality is available but often remains itself heterogeneous semantically and spatially, inaccessible, hermetic and in practice ends up to be neglected by most users. In fact, someone must develop a strong expertise to properly understand metadata and assess the fitness of given datasets and subsets for a specific use in well-defined areas and varying periods. Such a complex task involves thousands of partially correlated metadata. Consequently, data quality experts must rely on tools to help them pinpoint potential problems as well as synthesise the information necessary to write their opinion in a report involving their professional liability. In order to support such experts to assess fitness for use, this paper presents an approach aiming at better managing and communicating data quality information through a set of advanced database decision-support and visualisation concepts. This approach technically

relies on merging GIS capabilities with Business Intelligence technology (mostly On-Line Analytical Processing or OLAP), to adapt the executive dashboard approach and provide interactive, context-sensitive spatial data quality indicators. A prototype named MUM (Multidimensional User Manual) is presented to illustrate the approach.

5.3 Introduction

The last decade has witnessed a major trend towards the democratisation of geospatial data. These data are now used in various application domains and by a variety of users composed of people from experts with highly-sophisticated systems to mass-users with web and mobile mapping technologies. Although being a positive evolution, such democratisation also facilitates the use of data for non-intended purposes as well as the overlaying of heterogeneous data collected at different times by different organisations using various acquisition technologies, standards and specifications. Such context increases the risks of geospatial data misuse. In this sense, Goodchild (1995) argues that ‘GIS is its own worst enemy: by inviting people to find new uses for data, it also invites them to be irresponsible in their use’. Number of such cases already occurred, sometimes leading to significant social, political or economical impacts (e.g. Beard 1989; Monmonier 1994; Agumya and Hunter 1997; Gervais 2004).

In today's situation, it is difficult and sometimes impossible to clearly assess the fitness of certain data for a specific use over a given area. This is due, amongst others, to the inadequate documentation regarding data specifications, in spite of the development of standards over the past 10 years more particularly (e.g. FGDC, CEN, ISO, OpenGIS). An increasing number of papers were published in the last years to address the problem of the evaluation of fitness for use (e.g. Frank 1998; Agumya and Hunter 1999b; Agumya and Hunter 1999a; De Bruin *et al.* 2001; Vasseur *et al.* 2003; Grum and Vasseur 2004; Frank *et al.* Submitted). However, assessing the fitness for use is a very complex task and more research is needed to provide a simple and complete way to do it. On the legal side, as geospatial data can now be considered as a mass-product, one may argue they should follow the corresponding legislation and properly deal with consumer protection, liability, guarantees, clear instruction manuals, etc. In this context, data producers should be able to

communicate meaningfully quality information to users in order to help them assess the fitness of the data for their purpose (Gervais 2004).

Metadata (i.e. data about data) currently distributed by data producers should contribute to help assessing the fitness for use. However, metadata typically suffer from a large number of inter-related informations, a complex organisational structure, a high level of heterogeneity in their application, a lack of explicit links between metadata and data, an hermetic language, a highly complex content for both expert and non-expert users, a general lack of detail in their application, and so on. Hence, we can observe that currently, GIS users aren't able to get quality information that is easily accessible, understandable and adapted to their context and needs.

In order to support geospatial data users in the assessment of the fitness for use of their data, there is a need for improved methods and tools facilitating quality information management and communication. Such methods and tools would allow users to increase their knowledge about data quality and assess in which way data fit for their use. Several authors recently mentioned the need for such methods and tools. For instance, Lowell (2004) expresses the need for a 'computer-based intelligent engine' that could analyse information about uncertainty. He argues that 'Humans will not be able to absorb and assimilate all of the information presented in an uncertainty-based database, and will not have the capacity to analyse all of it efficiently. This will require the creation of new analytical and visualisation tools capable of providing humans with a logical summary of the uncertainty information present in the system'. Because of the complexity of this task, we think that it is currently impossible to design a system providing a clear output regarding the fit or un-fit of the data for a certain use. We argue that the only possibility available today and certainly in the near future is to provide users the required information regarding data quality and characteristics in order to help them making an 'informed decision' on the right data to use for a certain application in a given area. Furthermore, according to Gervais (2004), a non-expert user facing a complex assessment of fitness for use should request the opinion of an expert user or of an expert in geospatial data quality who will engage his professional liability into such an assessment and reduce the risk of misuse (cf. Figure 24). Consequently, the objective of this chapter is to present a Quality Information System called MUM (Multidimensional User

Manual) that aims to manage and communicate context-sensitive quality information to expert users and data quality experts.

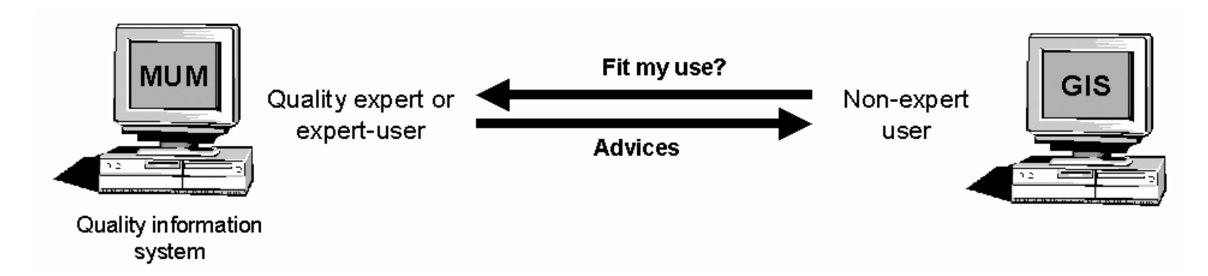


Figure 24: Quality Information System objective

We first discuss data quality management and communication. Then, we explain our approach that uses quality indicators based on quality information stored in a multidimensional data structure named Quality Information Management Model. We also discuss top-down and bottom-up approaches to populate the quality database. We then present our prototype system MUM which supports several techniques to manage and communicate quality information to the expert user or the data quality expert. The use of Spatial On-Line Analytical Processing (SOLAP) functions as well as the general architecture of the prototype are described, including its different functions and how it is used to support users during the quality assessment process. Finally, we discuss our results and conclude and the chapter with the proposal of future research directions.

5.4 Geospatial data quality management and communication

For about 30 years, two different meanings have been associated with the term ‘quality’ in the literature, the first one restricting quality to the absence of errors in the data (i.e. internal quality) and the second one looking at how data fit the user’s needs (i.e. external quality) (Juran *et al.* 1974; Morrison 1995; Aalders and Morrison 1998; Aalders 2002; Dassonville *et al.* 2002). This second definition, usually identified as the concept of ‘fitness for use’ (Juran *et al.* 1974; Chrisman 1983; Veregin 1999), is the one that reached an official agreement by standardisation bodies (e.g. ISO) and international organisations (e.g. IEEE). More precisely for the latter case, we define quality as the closeness of the agreement between data

characteristics and the explicit and/or implicit needs of a user for a given application in a given area.

For more than 20 years, standardisation bodies have identified characteristics describing internal quality (e.g. ICA, FGDC, CEN, ISO, OGC). If these characteristics differ between standards, there is however an agreement on most of them and common criteria are often identified as the 'famous five': positional accuracy, attribute accuracy, temporal accuracy, logical consistency and completeness (Guptill and Morrison 1995; ISO-TC/211 2002). It is intended to document these criteria within the metadata provided with datasets by data producers.

One objective of providing metadata is to allow end-users to assess the fitness of a dataset for their use (ISO-TC/211 2003). However, academic studies and practical experience clearly show the limited benefit of metadata in their current form (Timpf *et al.* 1996; Frank 1998; Gervais 2004). It is even not rare to see users asking producers not to give them metadata when ordering data. Our experience is that users rarely use metadata beyond the subset necessary for selecting and ordering datasets from digital libraries.

In addition to their inadequate form which is too hermetic for non-expert as well as several expert users, a strong limitation lies in the fact that metadata are often provided at a level of aggregation that is too general to enable an adequate quality assessment, hidding most of the information richness which should be communicated. Hunter (2001) clearly illustrates this point by giving several examples of existing metadata such as: Positional Accuracy being 'variable', '100m to 1000m' or '+/- 1.5m (urban) to +/- 250m (rural)'. Such metadata rapidly become useless when someone wants to know the quality of data for a certain region, object class or object instance for example.

Moreover, if metadata were not separated from data as it is curently done in most cases, quality information included within metadata could be directly exploited to enhance certain GIS functions. Let us consider for instance the simple case of a distance measurement between two objects on a map. A typical GIS will provide a very precise answer, whatever the data accuracy recorded in the metadata (e.g. ArcGIS 8.0 provides distances with six decimals, corresponding to a spatial precision of a thousandth of millimetre). Given the appropriate level of detail in metadata, it would however be possible to make the system get

the spatial data accuracy from the metadata and adapt the precision of the measurement according to it. Hence, from a more general point of view there is a real possibility of benefiting from the quality information described into metadata. The benefit would be twofold:

1. a more efficient communication of quality information would help users to assess how datasets fit for their use (i.e. an issue discussed in this paper);
2. the management of quality information into a structured database would allow, when associated with a GIS tool, to provide results adapted to the data manipulated for the area of interest (i.e. this is a research perspective).

Both points would help reducing the risk of misuse and then reduce the occurrence of adverse consequences.

During the past decade, several research projects have focused on ways to better communicate quality/uncertainty/error information, through for instance visualisation techniques (Buttenfield and Beard 1991; Beard and Mackaness 1993; Buttenfield 1993; McGranaghan 1993; Buttenfield and Beard 1994; Fisher 1994; Beard 1997; Beard and Buttenfield 1999; Leitner and Buttenfield 2000; Drecki 2002) or the communication of visual or audio warnings to users (Fisher 1994; Hunter and Reinke 2000; Reinke and Hunter 2002). However, none of these techniques is yet implemented into commercial GIS (although a few can easily be implemented within a GIS application). Furthermore, none of these techniques allows users to navigate intuitively into various categories of quality information, from one quality characteristic to another and from one level of detail to another. Finally, these approaches are not supported by an analytical data structure typical of modern decision-support technologies such as Dashboards, On-Line Analytical Processing (OLAP), datamarts and data mining, which are capable of managing, producing, analysing and communicating information at different levels of detail.

5.5 Quality indicators and Quality Information Management Model (QIMM)

5.5.1 Quality indicators

Since quality information can be described using different characteristics (e.g. accuracy, completeness, consistency, up-to-datedness), and since we are moving towards 'feature-level metadata', the volume of quality information increasingly becomes a problem when we try to efficiently communicate this information. In many domains people have to cope with the problem of meaningfully communicating large volumes of information in order to support decision-making processes. They often use 'indicators' that can be displayed into so-called 'dashboards' (also named 'balanced scorecards' or 'executive dashboards') to communicate relevant information to decision-makers (Kaplan and Norton 1992; Fernandez 2000; von Schirnding 2000; Goglin 2001). Based on traditional indicator-based methods, we adapted this approach for the geographic information context (cf. chapter 3). Indicators can be defined as 'a way of seeing the big picture by looking at a small piece of it' (Plan Canada 1999). Fernandez (2000) defines indicators as 'information or a group of information helping the decision-maker to appreciate a situation'. They indicate what is going on globally, allowing or not to go into the details. Let us take for instance a family doctor who wants to diagnose his patient's illness. The doctor knows that the human body is a complex system and that he cannot observe and measure all of its characteristics. Hence, he uses certain observations and measures (e.g. temperature, blood pressure, pulse) to get broad view of the patient's condition. In similar ways, number of organisations use indicators to assess what is going on in larger complex systems (e.g. economical indicators, social indicators or ecological indicators). Klein (1999) observed different types of decision-makers that have to make rapid decisions (e.g. firemen, aircraft pilots) and, based on these observations, he built the 'Recognition-Primed Decision model' that is well known in the decision-making community. He observed that indicators ('cues') are key components in decision-making processes and are used to characterise situations and choose which action to perform. Indicators are thought of as efficient synthetic key information about complex phenomena and provide global pictures and major trends. Typical strategic decision-making processes use a small number of indicators as one may see in numerous BI (Business Intelligence) applications and EIS

(Executive Information Systems). Typical indicators can be drilled down in a small number of layers that are expanded to provide available details when needed. Selecting the most relevant indicators among available ones or collecting new data to build a new indicator represents an interesting challenge when designing decision-support systems.

Using such indicators in a quality assessment decision-support system appears not only theoretically interesting, but realistically unavoidable in order to build a usable and credible system. With this in mind, context-sensitive quality information can be provided to the user at the right level of abstraction in order to help him identify quality aspects which are relevant for the task at hand. To analyse the fitness for use of geospatial data for a given area, we designed the MUM System such that quality indicators are displayed into a dashboard that is embedded within a cartographic interface, acting as a decision-support tool specific to data quality.

Each quality indicator can be based on a single raw data, or may be computed using several raw data. This data is obtained for instance from metadata provided with the datasets but can also be provided by other sources of information describing data quality such as an organisation's internal consensus about lower spatial precision for a given area or lower degree of completeness for a certain period within a dataset.

In the chapter 3, we identified two types of warnings that can be communicated to users: ‘manipulation warnings’ and ‘status warnings’. Manipulation warnings can warn users when a risk may occur from an incorrect data manipulation (as for example a risky combination of data and operator such as measuring the distance between a house and a parcel boundary when the latter is provided by an unofficial and imprecise source). Such issue was for instance discussed by Beard (1989) or Hunter and Reinke (2000). Status warnings provide information regarding the status of internal data quality. ‘Risk warnings‘ result from the comparison between internal data quality information and the user’s tolerance threshold (e.g. a data positional accuracy of 1 meter compared to a user threshold of 10 meters will result in an indicator that says that this aspect of quality is correct). They are expressed for instance on a qualitative ordinal scale, such as ‘exceed the needs’, ‘reach the needs’ or ‘below the needs’, which can be displayed using a green/yellow/red symbology. The qualification of such

quantitative quality data is a complex issue recently explored for geospatial data quality (see for instance Grum and Vasseur 2004; Frank *et al.* Submitted).

5.5.2 Quality Information Management Model (QIMM)

A central motivation in this research is to avoid an information overload to users, which can be caused by the various quality characteristics when described at different levels of detail. According to the well-known psychological research from Miller (1956), the short-term memory (or working-memory) of humans can only deal with five to nine chunks of information at once. Hence, it would be of limited use to communicate a large quantity of information simultaneously to a user. In addition, other psychological studies showed that the duration that information stays in short-term memory (STM) is very limited (Baddeley 1997). This duration can be quite variable depending on the modality (i.e. acoustic, visual or semantic), the necessity of performing actions (e.g. selecting an item on the screen of a computer) and other factors (for instance, the level of concentration). Experimental results usually provide durations varying from 2 to 30 seconds. According to Newell's (1990) physical and biological tests, among the four computational bands emerging from the natural hierarchy of information processing, respond times between 10^{-1} to 10^{1} seconds are needed to perform cognitive tasks and maintain a line of thoughts. Consequently, an efficient method to communicate quality information should limit the volume of information (less than nine chunks) and rapidly provide information to users in order to avoid interrupting his mind-stream. Another point highlighted by Reinke and Hunter (2002) is the need for users not only to get quality information from the system, but also to be able to interact with the system (i.e. feedback loop).

To cope with all these constraints, we base our approach on the multidimensional database model used in the field of Business Intelligence (data warehousing, OLAP, data mining). In this domain, ‘multidimensional’ does not refer to x, y, z and t as in the GIS domain but rather to semantic, temporal and spatial hierarchies of concepts called dimensions which are represented by the metaphor of a data hypercube containing facts; each fact containing measures resulting from the intersection of all dimensions at a given level in their hierarchy (see for instance Berson and Smith 1997). Multidimensional database approaches appeared in the early eighties (Rafanelli 2003) and numerous books and papers have been published on

this vast topic, especially after it became popular in the mid-nineties thanks to Codd (1993) who clearly explained the superiority of multidimensional databases over relational databases when the users need to interactively analyse large volumes of data. They now represent a very important aspect of decision-support database techniques, which were considered in the field of GIS only recently (see for instance Miller and Han 2001; Bédard *et al.* 2003).

Multidimensional databases are very well suited to facilitate quality analysis in data rich GIS applications since they are built especially to query data at different levels of granularity (avoiding information overload while allowing targeted drilling), to provide fast results from complex queries on large volumes of data (do not interrupt users' train-of-thought) and to allow an intuitive navigation into summarised or detailed interrelated information using different operators (providing interaction with the system).

In the chapter 4, we presented a model named QIMM allowing the management of quality information within a multidimensional database model. Quality information stored into the QIMM model is afterward manipulated using Spatial On-Line Analytical Processing (SOLAP; see Rivest *et al.* 2001; Bédard *et al.* 2003) to allow users to navigate into quality dimensions and to intersect them at any level of detail. The proposed model is based on two dimensions, namely 'Quality Indicator' and 'Analysed Data', both having 4 levels of granularity (cf. Figure 17). Users can explore quality information by navigating within the system at different levels of detail, going for instance along the 'Analysed Data' dimension to obtain the quality of an entire dataset down to the quality of a single object instance and even geometric primitive when available. In each case, the quality may refer to a global indicator down to a very specific characteristic of quality. Examples are presented later in this paper.

5.5.3 Populating the quality database: combining Bottom-up and Top-down approaches

Once a multidimensional database structure is designed to manage quality information, the next step is to fill this database with existing or derived quality information. Two approaches can be identified:

- *Bottom-up:* this approach aims at taking the quality information documented at detailed levels (e.g. spatial accuracy metadata for the geometric primitives of the National Topographic Database of Canada for instance) and to aggregate it into higher-level

information (e.g. average and standard deviation for the spatial accuracy of the ‘roads’ layer of the selected area, i.e. of all roads of this area).

- *Top-Down:* this approach consists in collecting more global quality information, such as an expert's opinion about the average spatial precision of planned roads in his county, and in propagating this general level information, when it is relevant, at detailed levels (e.g. each planned road of this county inheriting from his experts' opinion). For instance, it is typical to see land-surveyors having very good knowledge of a territory and of the quality of the different datasets describing it (e.g. cadastral and topographic data). Using their experience happens to frequently be the most reliable way to tell that a dataset is relevant or not for various applications in this area. They can also provide insights on the spatial heterogeneity of the quality of certain datasets, identifying higher and lower-quality regions in the area covered by the data. They can also do it with respect to the period of measurements and other informal criteria. New research has recently been undertaken by our research team to define how such implicit expert knowledge can be formalised and integrated into a quality management system.

If both approaches are complementary, they both have advantages and drawbacks. Indeed, in the first approach, metadata can be easier to collect, but finding the most efficient methods to aggregate quality information, to analyse and synthesise hundreds of metadata that vary over space and time can be a tricky issue. On the other hand, formalising expert opinions is not simple either, and the propagation of quality information to lower levels of details has to be done with caution because high-level information can be an implicit aggregation of heterogeneous low-level data. Nevertheless, it seems reasonable to believe that with today's knowledge, none of these approaches can completely fill the database, both could be used in most quality information systems, and the capacity of acquiring relevant data will be a key element when deciding which approach to choose. In addition, in the context of risk analysis for the use of data, one must keep in mind that ‘no information is information’ and ‘divergent information is also information’.

5.6 Applying the concepts: developing the Multidimensional User Manual (MUM) prototype

Based on the quality indicator approach and the QIMM structure, we developed a prototype software to support experts assessing the fitness of certain data for an intended use. The prototype implements, as a proof of concept, different operators which have been described in the chapter 3, such as displaying quality information using indicators, calculating indicators values according to the spatial extent visualised by the user, allowing users to select indicators relevant to their application, providing indicators at different levels of details, etc. In the next sections, we describe the architecture of this prototype, the quality indicators that make the multidimensional data structure and how experts can navigate into quality information.

5.6.1 Prototype architecture

The prototype was developed using four commercial off-the-shelf software driven by a unique user interface developed in Visual Basic (fast and easy for prototyping), which integrates the different mapping and database technologies (cf. Figure 25). These four main technologies are:

- *Microsoft SQL Server/Analysis Services*: this is the OLAP server that provides multidimensional database management functionalities with the MDX language;
- *Microsoft Access*: this popular relational database management system is used to store user profiles and multidimensional indicators' name and characteristics;
- *Proclarity*: this OLAP client software provides query and navigation functions (e.g. drill-down and roll-up operators) that allow users to explore the quality data stored into SQL Server;
- *Intergraph Geomedia Professional*: this Geographical Information System (GIS) software provides map-viewing functions such as *Zoom In*, *Zoom Out*, *Pan*, *Fit all* and other tools allowing the creation of quality maps.

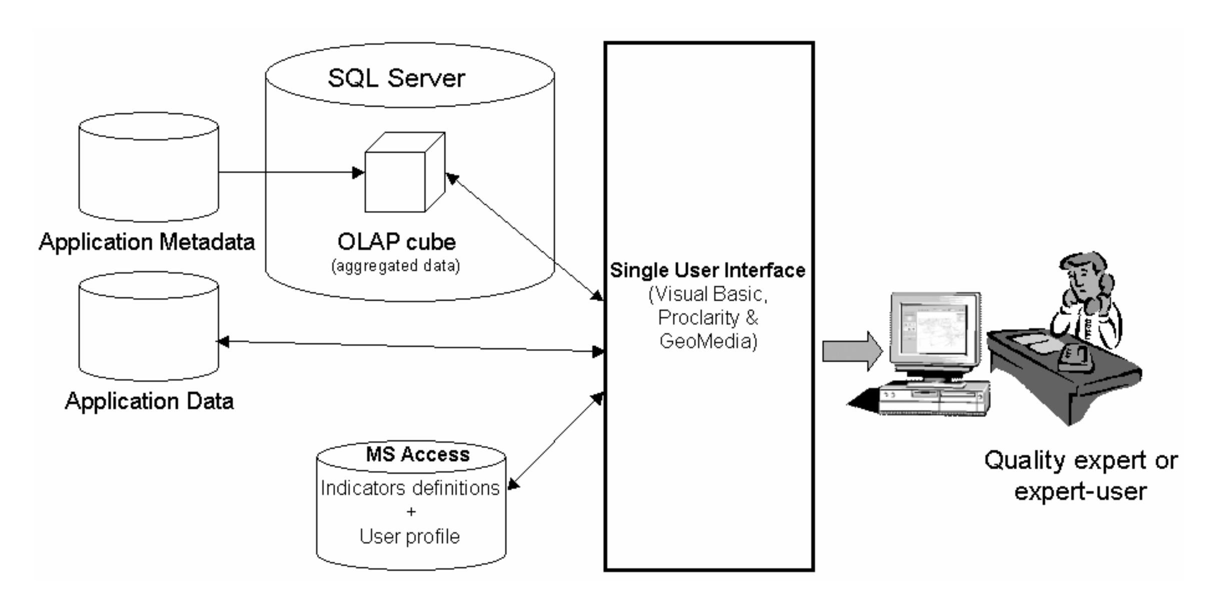


Figure 25: MUM prototype general architecture

Data quality information used for the experimentation was based on the recent ISO 19113 international standard dealing with quality information description (ISO-TC/211 2002). For increased speed, quality information is stored into the multidimensional database or data cube using a full Multidimensional OLAP data structure (MOLAP), as to the other possible relational OLAP structure (ROLAP) mimicking the former (see Berson and Smith 1997 for more details about the different OLAP architectures).

After a complete database design, making the proof of concept required to experiment with a subset of the QIMM dimensions within the prototype, including the entire indicator dimension and three levels of detail of the 'Analysed Data' dimension (i.e. dataset, data layer and object feature instance).

5.6.2 Indicators selection, calculation and representation

The quality indicator approach is based on the observation that (1) it is impossible in practice to obtain all detailed metadata and algorithmically derive a unique value for quality, (2) it is too complex to exhaustively consider all factors with their detailed spatial and temporal variability and (3) all users do not evaluate quality based on the same type of information. For instance, certain users will be more interested in spatial accuracy, others in data completeness. Certain persons will have an interest in temporal data quality aspects and others will not. For this reason, quality indicators can be selected by users according to their

needs. Based on the ISO 19113 standard, a set of quality indicators was defined and stored hierarchically into a relational database. Then, users can select the indicators they want to display in their analysis dashboard by simply applying a drag and drop operation from the indicator list to the dashboard creation tool (cf. Figure 26). Each indicator definition is stored within this database, including a description of what it represents, the way it is computed, some warnings related to its interpretation, its importance as defined by the user (expressed in term of weight), etc. The user can eventually adapt some items further. One may select among different graphical representations to illustrate each indicator (e.g. street light, smiley, speed meter).

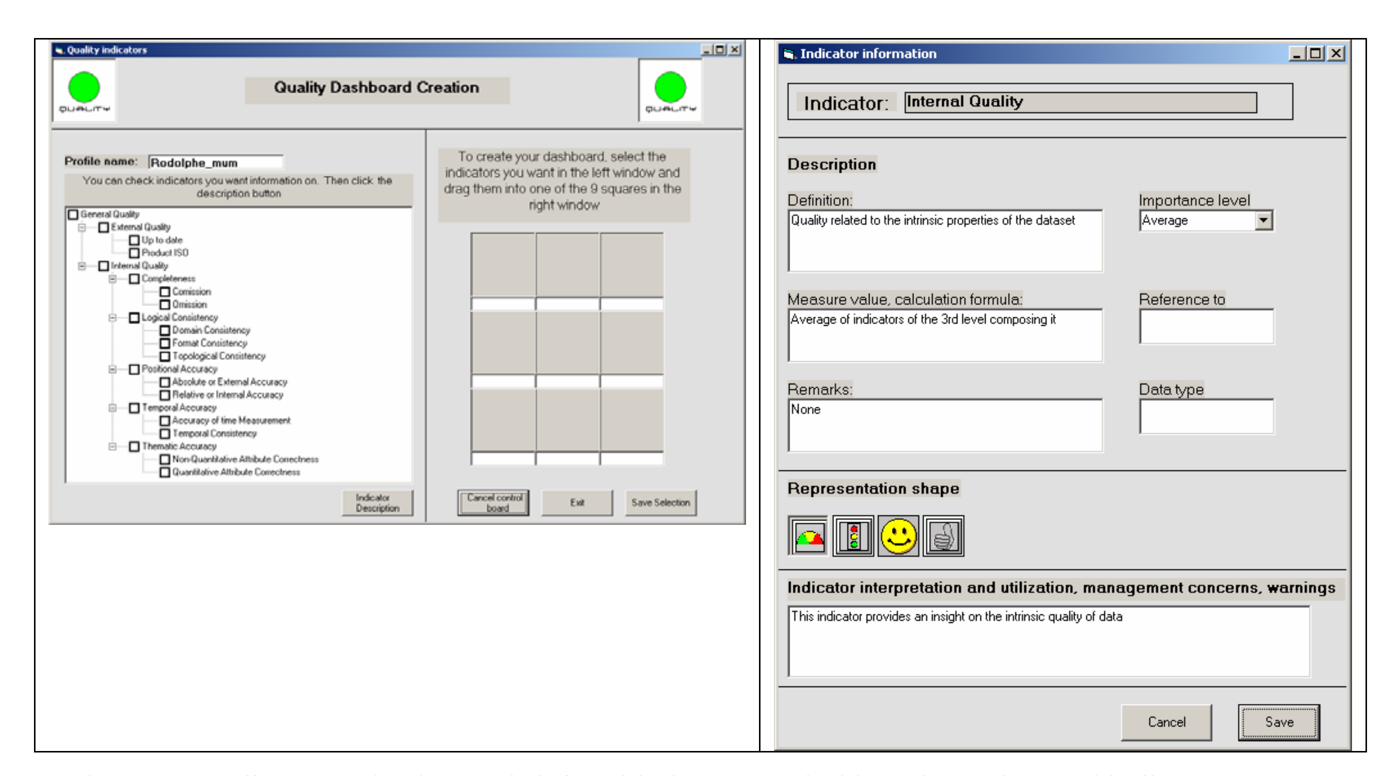


Figure 26: Indicators selection tool (left) with the empty dashboard template and indicators description and graphical representation form (right)

Indicator values are always based on the spatial extent visualised by the user. Indeed, if the user zoomed on a particular region of interest, it would not make sense to communicate quality information based on the objects located outside this area. Then, indicators' values are updated every time the user navigates into the map view using the '*zoom in*', '*zoom out*' or '*pan*' functions. *Ad hoc* polygon would also be of interest.

5.6.3 Navigation into Spatial Data Quality information

Using the prototype described in the previous section, geospatial data experts can improve their knowledge of data quality through the use of different navigation tools. Displaying information at different levels of detail within a short time period allows users to analyse the data quality without interrupting their line of thoughts. Figure 27 illustrates the benefits of such a system through different questions a user may have regarding data quality and the different tools offered by the system that can help answering these questions.

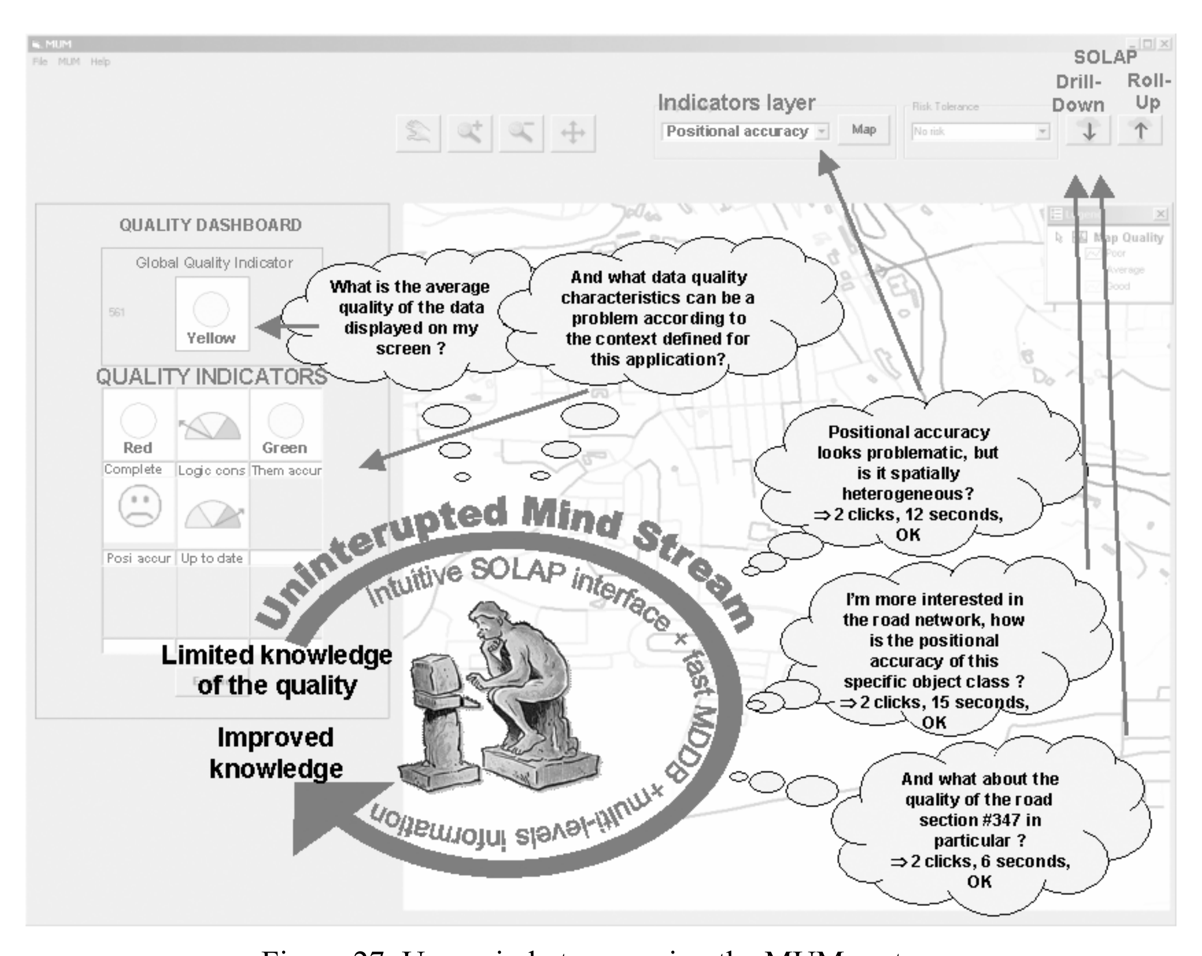


Figure 27: User mind-stream using the MUM system

Quality indicators (dashboard and global indicator)

Data quality information is communicated through the various indicators' possible representations (e.g. street light, smiley, speed meter) as well as quality maps. Using SOLAP operators, it is possible to drill on these representations as well as cartographic features. We provide a global indicator to represent the aggregation of all indicators for the displayed area.

Each indicator is the aggregation of sub-indicators, down to detailed metadata where it is possible. In our prototype, the quality dashboard can include up to nine indicators, which is consistent with Miller's rule (Miller 1956) that limits information volume to nine chunks for human short-term memory. The value of each quality indicator varies according to quality (e.g. an indicator using the street light representation can have the values green, yellow, red or white).

SOLAP navigation along the 'Analysed Data' dimension

SOLAP fast drill-down and roll-up capabilities are key elements of the prototype. They allow users to navigate from one level of detail to another along the 'Analysed Data' dimension. For instance, this allows users to get quality indicator values for the whole dataset, then look at the quality for a certain theme (e.g. only roads) and move again to get the quality of a single feature instance. Figure 28 illustrates this example of navigation. The prototype interface includes cartographic and SOLAP tools in the upper part, indicator dashboard including different indicators on the left side and the cartographic interface on the right side. These operators fully exploit the advantages of multidimensional databases, being intuitive and very fast.

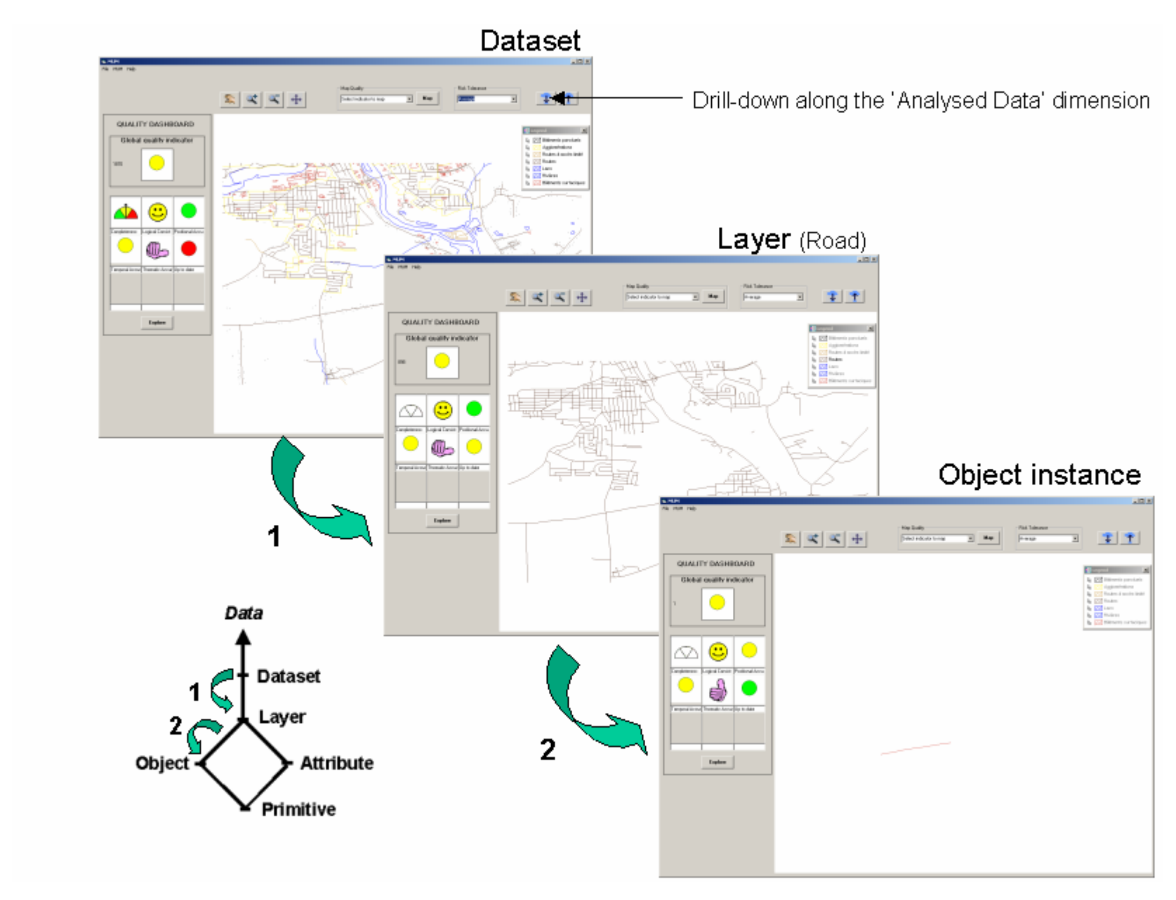


Figure 28: Navigation along the 'Analysed Data' dimension using two successive drill-down operations

SOLAP navigation along the 'Quality Indicator' dimension

Within the quality indicator dashboard, SOLAP drill-down and roll-up operators allow users to navigate from one level of detail to another along the 'Quality Indicator' dimension. Users can then explore quality indicators at the aggregated level and move down for instance to detailed levels when there seems to be a problem regarding quality (cf. Figure 29). Such an approach helps avoiding information overload and offers interactions between the user and the system. For instance, on the example of Figure 29, a user looks first at the higher-level indicators. He realises that 'General Quality' is only average (i.e. yellow) because of the lower 'Internal Quality'. He can then drill-down into the indicator hierarchy to see the sub-indicators composing the 'Internal Quality'. At this second level he can wonder why the 'Logical Consistency' indicator is only average and then drill-down again to get more details. He finally arrives at the last level of detail available and sees that the problem comes from the

'Topological Consistency'. He can then decide if this aspect of data quality is important for his application or not and then decide to either absorb the residual uncertainty of reduce it by, for instance looking for another dataset (Bédard 1987; Hunter 1999).

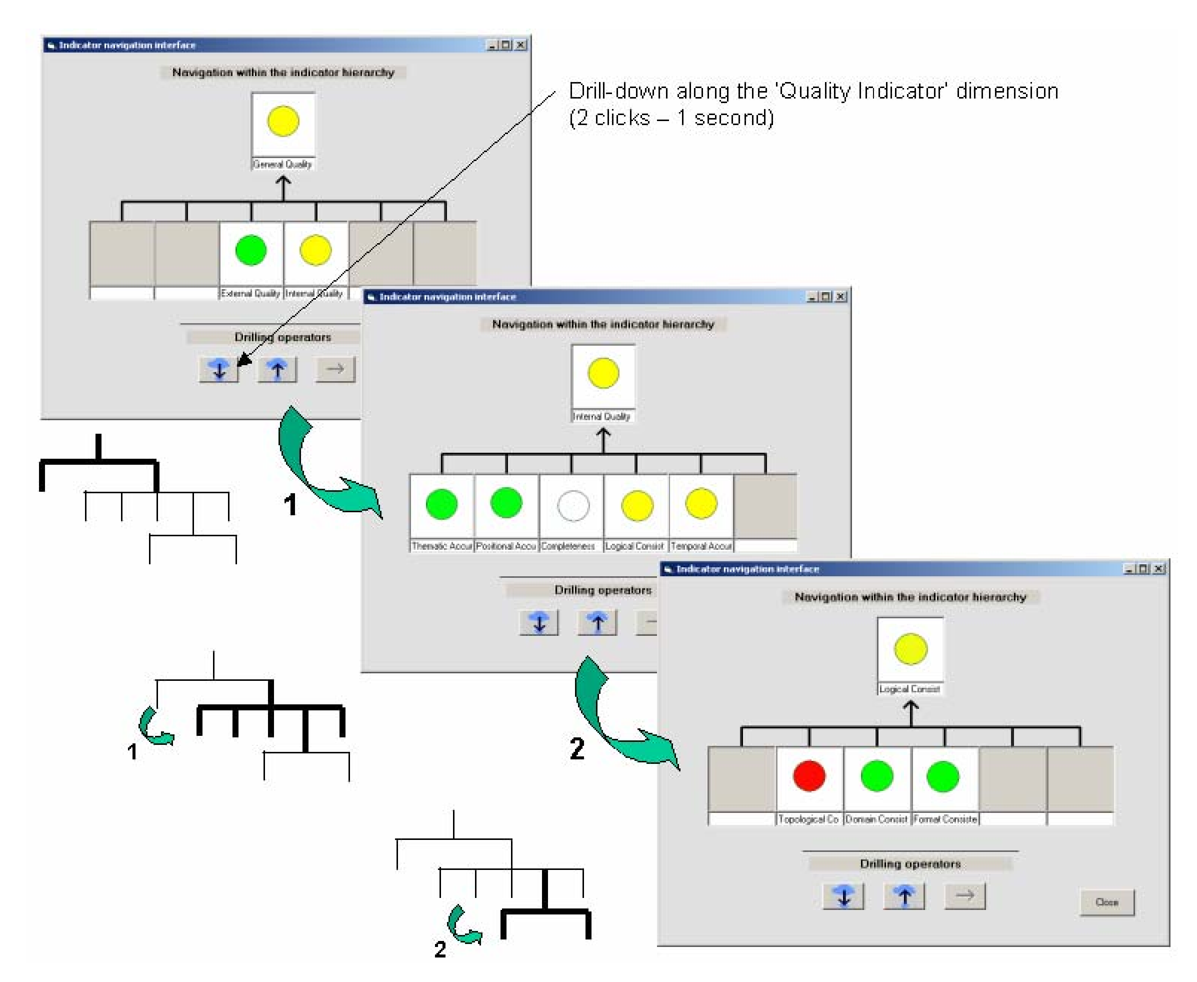


Figure 29: Navigation along the 'Quality Indicator' dimension using two successive drill-down operations

Indicator mapping

Indicator mapping allows users to get a fast insight on the spatial heterogeneity of a quality indicator. If metadata often document the average quality (e.g. spatial accuracy) for an entire map sheet, at a more detailed level quality can vary widely on a spatial basis. Let's take for instance a dataset covering a large area (e.g. country) that is the result of the integration of several datasets of various qualities that cover smaller adjacent areas (e.g. states). Without having such representation, the user could only get through metadata a unique quality value

and then underestimate or overestimate quality for specific areas. With our tool, users can explore quality through the indicators displayed in the dashboard. However, when drilling-down on the quality of each source, he could loose the global picture and quality analysis would then be more complicated. Indeed, it is difficult for users to get such a view from the indicators displayed into the dashboard. To get such an information, users would have to get quality indicators values successively for each feature instance. Quality mapping aims at tackle this issue.

Quality maps can use different types of classification according to the distribution of values. We implemented five different ways to create the qualitative classes: equal count, equal range, standard deviation, custom equal count and custom equal range. Changing the way to create classes can be useful, for instance, when all data of a certain dataset have similar quality levels. Instead of getting the same value (e.g. green) for all feature instances, it is then possible to highlight features having the lowest and the highest qualities in the distribution (cf. Figure 23).

5.7 Conclusion

This paper presented an approach helping expert users of geospatial data as well as data quality experts to improve their knowledge about data in order to assess their fitness for a given use. This approach is based on a multidimensional data structure (QIMM), that supports the fast and easy exploration of quality information at different levels of detail. Exploration goes along an ‘Analysed Data’ dimension as well as a ‘Quality Indicator’ dimension in addition to being supported by interactive quality mapping. Quality information is communicated to users through the contextual indicators displayed into a dashboard integrated into the SOLAP. The architecture of a prototype was presented as well as its main functionalities that allow users to navigate into diverse quality information at different levels of detail. This prototype was meant as a proof of the applicability of the proposed concepts, concepts which are considered the important results of this research. As such, the prototype only includes a subset of the possible functions that such a system could provide.

A validation of the approach was done through demonstrations of the prototype to different types of people from various domains (GIS scientists including specialists in data quality issues, consultants in GIS, data producers, governmental agencies, typical GIS users, etc.).

Such presentations of the project were performed since the early stages of the project in order to get an early feedback from potential users and then adapt the project in consequence. The different users expressed an interest in this approach and found it much more efficient than current metadata to increase users knowledge about data quality and then help to assess the fitness of data for certain use.

Different aspects of this research can be further explored in future research works, such as improving the model of user's needs/profile and formalise/integrate expert opinions into the QIMM model. Finally, it is worth mentioning that once quality information is stored in such a structured database with different levels of detail, quality information then becomes easily accessible and can be used to enhance many other aspects of a GIS application. This represents a step towards the creation of 'quality-aware GIS', which extends the concepts of Unwin's (1995) 'error-sensitive GIS' and of Duckham and McCreadie (2002) 'error-aware GIS'. We refer to a 'quality-aware GIS' as a GIS with the added capabilities to manage, update, explore, assess and communicate quality information. The term 'quality' encompassing more than 'error' by also addressing issues related to GIS users contexts and usages (e.g. user profile and needs assessment). This is then a step further towards better GIS.

Acknowledgement

This research is part of the MUM project (Multidimensional User Manual) and has benefited from financial support from the Canadian Network of Centres of Excellence GEOIDE, the IST/FET program of the European Community (through the REV!GIS project), the Ministère de la Recherche, de la Science et de la Technologie du Québec, the Canada NSERC Industrial Chair in Geospatial Databases for Decision-Support, the Centre for Research in Geomatics (CRG) and Université Laval. Thanks are due to Mathieu Lachapelle who contributed to the prototype development.

5.8 References

AALDERS, H.J.G.L., 2002, The Registration of Quality in a GIS. In *Spatial Data Quality*, edited by W. Shi, P. Fisher, and M.F. Goodchild, (Taylor & Francis), pp. 186-199.

AALDERS, H.J.G.L., and MORRISON, J., 1998, Spatial Data Quality for GIS. In *Geographic Information Research: Trans-Atlantic Perspectives*, edited by M. Craglia, and H. Onsrud, (London/Bristol: Taylor & Francis), pp. 463-475.

AGUMYA, A., and HUNTER, G.J., 1997, Determining fitness for use of geographic information. *ITC Journal*, **2**, 109-113.

AGUMYA, A., and HUNTER, G.J., 1999a, Assessing "fitness for use" of geographic information: What risk are we prepared to accept in our decisions ? In *Spatial Accuracy Assessment, Land Information Uncertainty in Natural Ressources*, edited by K. Lowell, and A. Jaton, (Quebec), pp. 35-43.

AGUMYA, A., and HUNTER, G.J., 1999b, A Risk-Based Approach to Assessing the 'Fitness for Use' of Spatial Data. *URISA Journal*, **11**, 33-44.

BADDELEY, A., 1997, *Human Memory: Theory and Practice* (U.K. Psychology Press).

BEARD, K., 1989, Use error: the neglected error component. In *Proceedings of AUTO-CARTO 9* (Baltimore, Maryland), pp. 808-817.

BEARD, K., 1997, Representations of Data Quality. In *Geographic Information Research: Bridging the Atlantic*, edited by M. Craglia, and H. Couclelis. (Taylor and Francis), pp. 280-294.

BEARD, K., and BUTTENFIELD, B., 1999, Detecting and evaluating errors by graphical methods. In *Geographical Information Systems*, edited by P.A. Longley, M.F. Goodchild, D.J. Maguire, and D.W. Rhind, (Wiley), pp. 219-233.

BEARD, K., and MACKANESS, W., 1993, Visual Access to Data Quality in Geographic Information Systems. *Cartographica*, **30**, 37-45.

BÉDARD, Y., 1987, Uncertainties in Land Information Systems Databases. In *Proceedings of Eighth International Symposium on Computer-Assisted Cartography* (Baltimore, Maryland), pp. 175-184.

BÉDARD, Y., GOSSELIN, P., RIVEST, S., PROULX, M.-J., NADEAU, M., LEBEL, G., and GAGNON, M.-F., 2003, Integrating GIS Components with Knowledge Discovery Technology for Environmental Health Decision Support. *International Journal of Medical Informatics*, **70**, 79-94.

BERSON, A., and SMITH, S.J., 1997, *Data Warehousing, Data Mining and OLAP (Data Warehousing / Data Management)* (McGraw-Hill).

BUTTENFIELD, B., and BEARD, K.M., 1994, Graphical and Geographical components of Data Quality. In *Visualization in Geographic Information Systems*, edited by H.M. Hearnshaw, and D.J. Unwin. (Wiley), pp. 150-157.

BUTTENFIELD, B.P., 1993, Representing Data Quality. *Cartographica*, **30**, 1-7.

BUTTENFIELD, B.P., and BEARD, K., 1991, Visualizing the quality of spatial information. In *Proceedings of AUTO-CARTO 10*, pp. 423-427.

CHRISMAN, N.R., 1983, The Role of Quality information in the Long Term Functioning of a Geographical Information System. In *Proceedings of International Symposium on Automated Cartography (Auto Carto 6)* (Ottawa, Canada), pp. 303-321.

CODD, E.F., 1993, *Providing OLAP (On-line Analytical Processing) to User-Analysts: An IT MandateReport*, E.F. Codd and Associates.

DASSONVILLE, L., VAUGLIN, F., JAKOBSSON, A., and LUZET, C., 2002, Quality Management, Data Quality and Users, Metadata for Geographical Information. In *Spatial Data Quality*, edited by W. Shi, P.F. Fisher, and M.F. Goodchild, (Taylor & Francis), pp. 202-215.

DE BRUIN, S., BREGT, A., and VAN DE VEN, M., 2001, Assessing fitness for use: the expected value of spatial data sets. *International Journal of Geographical Information Science*, **15**, 457-471.

DRECKI, I., 2002, Visualisation of Uncertainty in Geographic Data. In *Spatial Data Quality*, edited by W. Shi, P.F. Fisher, and M.F. Goodchild, (Taylor & Francis), pp. 140-159.

DUCKHAM, M., and MCCREADIE, J.E., 2002, Error-aware GIS Development. In *Spatial Data Quality*, edited by W. Shi, P.F. Fisher, and M.F. Goodchild, (London: Taylor & Francis), pp. 63-75.

FERNANDEZ, A., 2000, *Les nouveaux tableaux de bord des décideurs* (Éditions d'organisation).

FISHER, P., 1994, Animation and sound for the visualization of uncertain spatial information. In *Visualization in Geographic Information Systems*, edited by H.M. Hearnshaw, and D.J. Unwin, (Wiley), pp. 181-185.

FRANK, A.U., 1998, Metamodels for Data Quality Description. In *Data Quality in Geographic Information - From Error to Uncertainty*, edited by M.F. Goodchild, and R. Jeansoulin, (Editions Hermes), pp. 15-29.

FRANK, A.U., GRUM, E., and VASSEUR, B., Submitted, How to select the Best Dataset for a Task? *International Journal of Geographical Information Science*.

GERVAIS, M., 2004, Pertinence d'un manuel d'instructions au sein d'une stratégie de gestion du risque juridique découlant de la fourniture de données géographiques numériques. Ph.D. thesis, Sciences Géomatiques, Université Laval, Québec.

GOGLIN, J.-F., 2001, *Le datawarehouse pivot de la relation client* (Hermès Sciences).

GOODCHILD, M.F., 1995, Sharing Imperfect Data. In *Sharing Geographic Information*, edited by H.J. Onsrud, and G. Rushton, (New Brunswick, NJ: Rutgers University Press), pp. 413-425.

GRUM, E., and VASSEUR, B., 2004, How to select the best dataset for a task? In *Proceedings of 3rd International Symposium on Spatial Data Quality (ISSDQ'04)* (Bruck an der Leitha, Austria), pp. 197-206.

GUPTILL, S.C., and MORRISON, J.L., 1995, *Elements of spatial data quality* (Elsevier Science).

HUNTER, G.J., 1999, Managing uncertainty in GIS. In *Geographical Information Systems*, edited by P.A. Longley, M.F. Goodchild, D.J. Maguire, and D.W. Rhind, (John Wiley & Sons, Inc.), pp. 633-641.

HUNTER, G.J., 2001, Spatial Data Quality Revisited. In *Proceedings of GeoInfo 2001* (Rio de Janeiro, Brazil), pp. 1-7.

HUNTER, G.J., and REINKE, K.J., 2000, Adapting Spatial Databases to Reduce Information Misuse Through Illogical Operations. In *Proceedings of 4th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences (Accuracy 2000)* (Amsterdam), pp. 313-319.

ISO-TC/211, 2002, *Geographic Information - Quality principles, Report*, 19113.

ISO-TC/211, 2003, *Geographic Information - Metadata, Report*, 19115.

JURAN, J.M., GRYNA, F.M.J., and BINGHAM, R.S., 1974, *Quality Control Handbook* (McGraw-Hill).

KAPLAN, R., and NORTON, D., 1992, The balanced scorecard: Measures that Drive Performance. *Harvard Business Review*, **70**, 71-79.

KLEIN, G., 1999, *Sources of Power - How people make decisions* (MIT Press).

LEITNER, M., and BUTTENFIELD, B.P., 2000, Guidelines for the Display of Attribute Certainty. *Cartography and Geographic Information Science*, **27**, 3-14.

LOWELL, K., 2004, Why aren't we making better use of uncertainty information in decision-making? In *Proceedings of 6th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences* (Portland, Maine, USA).

MCGRANAGHAN, M., 1993, A cartographic View of Spatial Data Quality. *Cartographica*, **30**, 8-19.

MILLER, G.A., 1956, The Magical Number Seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, **63**, 81-97.

MILLER, H.J., and HAN, J., 2001, *Geographic Data Mining and Knowledge Discovery* (Taylor & Francis).

MONMONIER, M., 1994, A Case Study in the Misuse of GIS: Siting a Low-Level Radioactive Waste Disposal Facility in New-York State. In *Proceedings of Conference on Law and Information Policy for Spatial Databases* (Tempe (AZ) USA), pp. 293-303.

MORRISON, J.L., 1995, Spatial data quality. In *Elements of spatial data quality*, edited by S.C. Guptill, and J.L. Morrison, (New York: Elsevier Science inc.).

NEWELL, A., 1990, *Unified theories of cognition* (Harvard University Press).

PLAN CANADA, 1999, *Sustainable community indicators program, Report*, Vol 39 (5).

RAFANELLI, M., 2003, *Multidimensional Databases: Problems and Solutions* (Idea Group Publishing).

REINKE, K.J., and HUNTER, G.J., 2002, A Theory for Communicating Uncertainty in Spatial Databases. In *Spatial Data Quality*, edited by W. Shi, P.F. Fisher, and M.F. Goodchild, (London: Taylor & Francis), pp. 77-101.

RIVEST, S., BÉDARD, Y., and MARCHAND, P., 2001, Towards Better Support for Spatial Decision Making: Defining the Characteristics of Spatial On-Line Analytical Processing (SOLAP). *Geomatica*, **55**, 539-555.

TIMPF, S., RAUBAL, M., and KUHN, W., 1996, Experiences with Metadata. In *Proceedings of Symposium on Spatial Data Handling, SDH'96, Advances in GIS Research II* (Delft, The Netherlands), pp. 12B.31 - 12B.43.

UNWIN, D., 1995, Geographical information systems and the problem of error and uncertainty. *Progress in Human Geography*, **19**, 549-558.

VASSEUR, B., DEVILLERS, R., and JEANSOULIN, R., 2003, Ontological approach of the fitness of geospatial datasets. In *Proceedings of 6th Agile Conference on Geographic Information Science* (Lyon, France), pp. 497-504.

VEREGIN, H., 1999, Data quality parameters. In *Geographical Information Systems*, edited by P.A. Longley, M.F. Goodchild, D.J. Maguire, and D.W. Rhind, (John Wiley & Sons, Inc.), pp. 177-189.

VON SCHIRNDING, Y.E., 2000, Health-and-environment indicators in the context of sustainable development. In *Proceedings of Consensus Conference on Environmental Health Surveillance: Agreeing on basic set of indicators and their future use* (Quebec city, Canada).

Chapitre 6 : Conclusion

6.1 Sommaire

Cette thèse a présenté une approche visant à gérer, communiquer et faciliter l'analyse rapide de l'information sur la qualité des données géospatiales.

Le chapitre 1 a introduit les contexte et problématique abordés par la thèse, à savoir le besoin d'outils permettant de communiquer et analyser l'information sur la qualité des données géospatiales afin de supporter les utilisateurs dans l'évaluation de l'adéquation des données à leur utilisation (*fitness for use*).

Le chapitre 2 a présenté une revue de littérature autour des concepts reliés à la thèse. Nous avons abordé dans un premier temps l'incertitude existant dans les systèmes d'information géographique et la place de cette incertitude dans les processus de prise de décision utilisant des SIG. Nous avons ensuite présenté la terminologie reliée à la qualité, le concept de qualité en tant que tel puis, plus spécifiquement, le concept de qualité pour les données géospatiales. Nous avons dans un troisième temps décrit le processus d'évaluation de la qualité ainsi que différents travaux ayant porté sur la gestion et la communication des informations sur la qualité.

Le chapitre 3 a présenté les concepts d'indicateur et de tableaux de bord de qualité, présentant des informations sur la qualité des données aux utilisateurs du système. Les indicateurs permettent d'appréhender l'information sur la qualité des données de manière contextuelle par le biais d'une sélection des indicateurs dans une base de données d'indicateurs prédéfinis. Les indicateurs, organisés hiérarchiquement, sont affichés dans un tableau de bord intégré à l'interface du SIG, à la manière des SOLAP, évitant ainsi de communiquer un volume trop important d'information aux utilisateurs. Les principales caractéristiques que devrait posséder le système utilisant les tableaux de bord sont identifiées. Les indicateurs sont présentés comme un type d'avertissement fait aux utilisateurs et une classification de ces avertissements est proposée identifiant deux types d'indicateurs, soit les indicateurs de statut et de risque.

Le chapitre 4 a présenté un modèle multidimensionnel permettant la gestion des données sur la qualité à différents niveaux de détails. Le problème de la granularité de l'information sur la qualité ainsi que de la diversité des caractéristiques possibles pouvant décrire la qualité sont tout d'abord abordés et les principaux travaux existants portant sur ces aspects sont présentés. Les bases de données multidimensionnelles et les approches SOLAP sont par la suite introduites, puis le modèle QIMM permettant une gestion multidimensionnelle de l'information sur la qualité est présenté. La flexibilité et la richesse du modèle QIMM sont illustrées par des exemples de navigation possibles à l'intérieur des informations stockées sur la qualité. Enfin, des exemples de visualisation possible de la qualité, basés sur le modèle QIMM, sont présentés.

Finalement, le chapitre 5 a présenté un prototype fonctionnel du système MUM. Cet outil est basé sur le modèle QIMM pour la gestion des informations sur la qualité, permet leur communication sous la forme d'indicateur et de visualisation cartographique et leur analyse grâce à des opérateurs d'analyse multidimensionnelle permettant d'explorer la qualité. Des données extraites de la Base Nationale de Données Topographiques du Canada (BNDT) et des métadonnées suivant la norme ISO 19113 ont été utilisées pour l'implémentation du prototype. L'architecture du système est présentée à travers les différentes composantes logicielles ainsi que les processus que suivent les données, des données brutes aux données agrégées. Par la suite, différentes fonctions du prototype sont présentées et illustrées, montrant l'utilisation intuitive et rapide pouvant être faite de ce type d'outils.

6.2 Discussion

Cette thèse a présenté une approche visant à gérer, communiquer et faciliter l'analyse rapide de l'information sur la qualité des données géospatiales. Cette approche permet de communiquer à des usagers experts différentes caractéristiques de la qualité sous la forme d'indicateurs qui sont affichés dans un tableau de bord ou représentés sur une base cartographique. L'usager peut sélectionner les indicateurs dont il a besoin parmi un ensemble d'indicateurs disponibles, choisir un type de représentation et définir un niveau de risque qu'il est prêt à prendre. Le système lui offre différents opérateurs lui permettant de naviguer dans ces informations à différents niveaux de détails. Une représentation cartographique des

indicateurs est également proposée et permet de mieux appréhender l'hétérogénéité spatiale de la qualité.

L'objectif principal de la thèse était de « proposer une nouvelle approche permettant de gérer des données décrivant la qualité des données qu'un usager manipule et de les diffuser sous une forme plus compréhensible à des usagers experts ou des experts en qualité de données géospatiales ». Cet objectif a donc été atteint. Les deux sous-objectifs de la thèse ont été atteints et les travaux ayant permis de les atteindre sont présentés dans les chapitre 3, puis 4 et 5 respectivement.

Notre hypothèse de départ était : « il est possible de fournir aux utilisateurs experts ou aux experts en qualité des indicateurs renseignant sur les différentes caractéristiques de la qualité. Ces indicateurs de qualité peuvent être communiqués de manière contextuelle et à différents niveaux de détails et être intégrés dans un système plus large permettant de supporter les experts dans l'évaluation de l'adéquation des données à une utilisation. ». Nous pensons donc que l'hypothèse de départ a été vérifiée.

Une validation de notre approche a été effectuée en présentant les concepts et le prototype à divers intervenants, experts en géomatique ou non, scientifiques, industriels, représentants du gouvernement, etc. Ces présentations ont été faites à différents stades du projet (de l'idée initiale jusqu'au prototype final). Des utilisateurs ont été amenés à utiliser le prototype, ce qui a aidé à améliorer l'interface, identifier de nouveaux besoins et constater l'intérêt de cette approche. Ces démonstrations ont ainsi permis de mieux orienter la recherche en fonction des besoins de la communauté. Cette validation a permis de constater que les intervenants ont trouvé l'approche intéressante et beaucoup plus utile que les métadonnées actuellement fournies. Un représentant d'une organisation produisant des données géospatiales a également trouvé un intérêt dans cette approche comme un outil pouvant faciliter la planification de la production de leurs données (ex. identification visuelle rapide de la qualité des données permettant une planification des mises à jour en donnant priorité aux zones de moins bonne qualité). Il aurait été intéressant d'étendre cette validation en intégrant différents jeux de données et en comparant l'utilisation des données faite avec et sans le système. Toutefois, une telle approche aurait nécessité des temps de développement, et donc financiers, dépassant largement le cadre de cette thèse.

Le modèle QIMM présenté dans cette thèse permet une modélisation plus poussée des informations sur la qualité que les solutions proposées par d'autres auteurs (ex. Qiu et Hunter, 1999 et 2002; Faïz, 1999). En effet, en plus de descendre à un niveau plus détaillé dans la dimension des données (en allant aussi gérer la sémantique), le modèle permet aussi de hiérarchiser les indicateurs de qualité pour alléger le volume d'informations communiquées en même temps aux usagers. La structure de données de type OLAP permet de plus une exploitation plus efficace de l'information sur la qualité que les structures de données traditionnelles (ex. relationnelles). L'adaptation des approches de tableaux de bord de gestion pour communiquer les informations sur la qualité n'a pas de précédent dans la littérature. Si certains auteurs utilisent le terme « indicateur » de qualité, ce n'est pas toutefois dans la même optique, les indicateurs de gestion étant contextuels aux utilisateurs. Cette thèse est également la première à utiliser des outils de type OLAP (et SOLAP) pour gérer et communiquer les informations sur la qualité de données géospatiales, permettant une communication dynamique des informations sur la qualité.

Cette thèse fait partie d'un projet plus large nommé MUM dans lequel s'inscrivent la thèse du Dr. Marc Gervais (2004) et le mémoire de M. Johan Lévesque (début 01/2005). Elle n'offre donc pas toutes les solutions aux problèmes traités par MUM. Ainsi, quoique au début ce projet visait aussi les utilisateurs non-experts, il a évolué pour s'intéresser (suite aux résultats de la thèse de M. Gervais) spécifiquement aux utilisateurs experts. Certains outils pourraient probablement être mis à la disposition des deux types d'utilisateurs, mais nous pensons que le manque de connaissance en information géographique des utilisateurs non-experts ne permet pas de communiquer le même type d'information. D'autres recherches seront nécessaires afin d'identifier les moyens les plus appropriés de leur communiquer l'information sur la qualité. D'autres éléments de discussion sont proposés dans la section « perspectives de recherche ».

6.3 Conclusions

Cette thèse permet de tirer différentes conclusions :

- Il nous apparaît possible de mettre au point des outils efficaces et intuitifs permettant à des utilisateurs experts ou des experts en qualité, d'analyser la qualité de données géospatiales. Ce type de système permet à ces utilisateurs d'accroître leur connaissance

de la qualité et d'être ainsi à même de mieux appréhender des risques potentiels pouvant émerger de l'utilisation de données de qualité inappropriée;

- Les métadonnées ont dans leur forme et mode de transmission actuels de nombreuses limitations. En effet, en plus d'être rarement transmises aux utilisateurs, de ne pas être lues par ces derniers (i.e. mode de communication inapproprié), elles sont généralement incomplètes (ne décrivant que certains aspects de la qualité), sont présentées à un niveau trop général, ne sont pas reliées aux données (pouvant ainsi créer des problèmes de mise à jour), etc. De plus, leur format, souvent textuel, n'est ni facilement exploitable automatiquement par des systèmes informatiques, ni facilement compréhensible par des utilisateurs. Toutefois, malgré ces limites, les métadonnées sont plus que jamais nécessaires, comme données sources, pour permettre une communication plus compréhensible des informations sur la qualité sous une autre forme;

- Étant donné les limites des métadonnées que l'on peut observer, les métadonnées ne devraient pas être le produit final transmis aux utilisateurs, mais un produit intermédiaire, intimement lié aux données, pouvant être exploité par des systèmes informatisés qui pourront communiquer plus clairement les informations sur la qualité. Pour ce faire, les métadonnées fournies par les producteurs devraient suivre des normes (ex. ISO 19115) et être formalisées le plus possible (ex. éviter les descriptions faites sous la forme de texte libre) afin d'en faciliter le traitement en fonction d'une présentation finale, par exemple sous la forme de cube. Elles devraient décrire les données à différents niveaux de détails afin de permettre une communication plus précise et donc plus riche des informations sur la qualité;

- La technologie SOLAP ouvre de nouvelles possibilités pour la gestion et l'exploration des données de qualité. Les bases de données multidimensionnelles sont en effet adaptées à la gestion d'informations sur la qualité, celles-ci pouvant être documentées à différents niveaux de détails. Les opérateurs de type SOLAP (ex. *drill-down spatial, roll-up thématique*) permettent de naviguer intuitivement dans l'information sur la qualité tout en évitant une surcharge d'information. De plus, les performances offertes par les outils de type SOLAP rencontrent des critères cognitifs en terme de temps de réponse des différents opérateurs;

- Les indicateurs peuvent être avantageusement adaptés au domaine de la géomatique comme outils de support à la prise de décision. Ces outils peuvent être intégrés dans des logiciels de cartographe existants (e.g. SIG, SOLAP), peuvent être adaptés en fonction des besoins et apportent une solution intéressante pour communiquer de larges volumes de métadonnées sans surcharger l'utilisateur d'information;

- Le MUM communique l'information sur la qualité de manière plus efficace et plus intuitive que les métadonnées traditionnelles. Il offre entre autres une visualisation spatiale de la qualité permettant de mieux caractériser l'hétérogénéité de la qualité. Cette prise en compte de l'hétérogénéité spatiale devrait gagner en importance dans les années à venir. En effet, les données manipulées par des utilisateurs tendent à (1) résulter de plus en plus de la fusion de données provenant de différentes sources hétérogènes et (2) les processus de mise à jour risquent de plus en plus de passer d'un fonctionnement où on mettait à jour l'ensemble des objets d'un feuillet cartographique, à des mises à jour par occurrence et par classe d'objets ayant changées sur le territoire. Ces deux changements dans le processus de production vont résulter en des jeux de données de qualité très hétérogène.

- Le prototype développé a reçu un accueil très favorable là où il a été présenté et nous porte à croire que l'approche proposée constitue bel et bien une solution non seulement novatrice mais également une solution qui possède un fort potentiel d'applicabilité.

6.4 Perspectives de recherche

L'approche présentée dans cette thèse pour la gestion et la communication de l'information sur la qualité ouvre de nouvelles perspectives pour l'élaboration de logiciels de cartographie plus sensibles aux problèmes de qualité. Toutefois, certains aspects mériteraient d'être explorés ou approfondis afin d'améliorer cette approche :

- L'intégration des métadonnées et des données pourrait être automatisée pour permettre une analyse rapide de nouveaux jeux de données dans la base de données multidimensionnelle. L'utilisation d'un format tel que XML pourrait alors être explorée (ex. tel qu'utilisé dans le logiciel ArcGIS). De plus, des correspondances entre différentes normes de métadonnées (c.à.d. *crosswalks*, comme celles supportées par M^3Cat de la

compagnie Intélec de Montréal) pourraient être implantées pour permettre l'intégration automatique de métadonnées structurées selon ces différentes normes. Les correspondances entre les normes doivent alors être rigoureusement établies afin d'éviter des confusions entre métadonnées identiques portant des noms différents dans différentes normes, ainsi que le cas inverse;

- Le calcul des indicateurs de risque résultant de la comparaison entre les données décrivant la qualité (c.à.d. métadonnées) et les besoins des utilisateurs, est un problème complexe faisant l'objet, en géomatique, de récentes études (Grum et Vasseur, 2004; Frank *et al.*, Soumis). Les métriques utilisées dans cette thèse sont empiriques, comme le sont les autres méthodes citées dans la littérature, mais pourraient être raffinées pour être spécifiées de façon à mieux tenir compte du contexte des utilisateurs. Une approche par logique floue pourrait être explorée afin de mieux nuancer les limites « floues » séparant des données « acceptables » de données « inacceptables »;
- La qualité résulte de la comparaison entre les différentes caractéristiques des données et les besoins des utilisateurs. Les besoins sont exprimés à travers la sélection contextuelle des indicateurs et la proposition de différentes méthodes d'agrégation des métadonnées, suivant le niveau de risque accepté par l'utilisateur. Toutefois, le processus de définition des besoins pourrait être beaucoup plus approfondi. Les approches développées dans le domaine du *User Modeling* en intelligence artificielle pourraient entre autres être explorées (Fisher, 2001; Kobsa, 2001). L'intégration d'une approche ontologique pour la formalisation des besoins et des caractéristiques des jeux de données, telle que développée dans le projet REVIGIS, pourrait également être explorée;
- L'approche présentée dans cette thèse agrège des métadonnées pour en déduire des indicateurs (c.à.d. approche *bottom-up*). Les données sur la qualité pourraient à l'inverse, et de façon complémentaire, être documentées à un niveau plus général par des experts, puis être propagées à des niveaux de détails plus fins (c.à.d. approche *top-down*). Cette approche est brièvement présentée dans le chapitre 5 mais n'a pas été implémentée dans le prototype MUM. Cette approche pourrait permettre, entre autres, de remédier aux cas où il y a peu de métadonnées disponibles;

- Il serait intéressant d'effectuer une validation plus poussée de l'approche afin de mieux qualifier le bénéfice qu'offre une telle approche en comparaison aux approches actuellement disponibles (ex. diffusion simple de métadonnées). Une telle validation, pour être pertinente, aurait nécessité d'être faite en « grandeur réelle », c.à.d. dans un contexte réel d'utilisation (i.e. un ou plusieurs projets), avec un nombre significatif d'utilisateurs et pour différentes applications intégrant différents jeux de données. Cela permettrait par exemple de comparer l'utilisation des données faite avec et sans le système. Une telle approche aurait cependant nécessité des temps de développement, et donc financiers, dépassant largement le cadre de cette thèse, mais pourrait être effectuée dans le cadre de développements futurs (un mémoire de MSc débutant en janvier 2005 devrait porter sur cet aspect avec des données du ministère des Transports du Québec);

- Les méthodes/outils développés dans cette thèse visent des utilisateurs experts ou des experts en qualité. Il existe cependant un besoin réel pour rendre ce type d'approche accessible à des utilisateurs non-experts. Toutefois, l'ensemble des fonctionnalités offertes pour les experts peuvent ne pas convenir à des non-experts. Des travaux futurs pourraient évaluer dans quelle mesure cette approche peut être adaptée à des utilisateurs non-experts, basé sur des considérations à la fois légales et technologiques, mais aussi en terme d'efficacité du processus de communication;

- L'approche présentée dans cette thèse permet de communiquer plus efficacement des informations sur la qualité à des experts qui vont alors pouvoir mieux conseiller d'autres utilisateurs non-experts en qualité de l'information géographique. Toutefois, le lien existant entre la qualité des données et la qualité de la décision faite pourra être approfondi. C'est-à-dire voir dans quelle mesure certains problèmes de qualité, d'amplitude variable, auront des impacts sur les décisions qui vont être prises. Certaines équipes de recherche telle que S. de Bruin (Pays-Bas) et G. Hunter (Australie) s'intéressent à ces problèmes. Une exploration plus poussée de ces aspects pourra permettre de mieux cerner l'impact qu'aura une telle approche sur la communauté d'utilisateurs finaux.

- L'approche présentée dans cette thèse répond à certaines composantes que différents auteurs appellent « *Error-Aware GIS* », « *Quality-Aware GIS* » ou encore « *Intelligent*

GIS » (Burrough, 1992; Unwin, 1995; Duckham and McCreadie, 2002). Elle offre, entre autres, une méthode permettant de gérer, communiquer et analyser l'information sur la qualité. D'autres fonctionnalités pourraient être ajoutées au système MUM telles que des outils permettant de mettre à jour les métadonnées lorsque des changements sont effectués sur les données, des techniques de propagation d'incertitude permettant d'évaluer l'incertitude résultante lors de certaines manipulations, etc. Il serait également intéressant d'explorer les façons dont l'information sur la qualité pourrait être exploitée de façon plus systématique par les fonctions des SIG afin de prendre en compte automatiquement la qualité lors des opérations faites avec un SIG (ex. mesure de distance, calcul de nombre d'entités présentes dans une zone).

Comme mentionné précédemment, l'approche présentée dans cette thèse pour la gestion et la communication de l'information sur la qualité ouvre de nouvelles perspectives de recherche, tant théoriques qu'applicatives. Elle aura constitué, nous le souhaitons, une contribution d'intérêt pour la communauté intéressée à la qualité des données géospatiales.

6.5 Références

Burrough P. A., "Development of intelligent geographical information systems", *International Journal of Geographical Information Systems*, vol. 6, nº 1, 1992, p. 1-11.

Duckham M., McCreadie J. E., "Error-aware GIS Development". *Spatial Data Quality* (W. Shi, P.F. Fisher, and M.F. Goodchild, Eds), Taylor & Francis, London, UK, p. 63-75, 2002.

Faïz, S.O., 1999. "Systèmes d'Informations Géographiques: Information Qualité et Data Mining", Tunis, Éditions C.L.E., 362 p.

Fisher G., "User Modeling in Human-Computer Interaction", *User Modeling and User-Adapted Interaction*, vol. 11, 2001, p. 65-86.

Frank A. U. Grum E., Vasseur B., "How to select the Best Dataset for a Task?" *International Journal of Geographical Information Science*, Soumis.

Grum E., Vasseur B., "How to select the best dataset for a task?" *Proceedings of 3rd International Symposium on Spatial Data Quality (ISSDQ'04)*, Bruck an der Leitha, Autriche, 15-17 avril 2004, p. 197-206.

Kobsa A., "Generic User Modeling Systems", *User Modeling and User-Adapted Interaction*, vol. 11, 2001, p. 49-63.

Qiu, J., and G.J. Hunter, 1999. "Managing Data Quality Information", Proceedings of International Symposium on Spatial Data Quality, 18-20 juillet 1999, Hong Kong, p. 384-395.

Qiu, J., and G.J. Hunter, 2002. "A GIS with the Capacity for Managing Data Quality Information". Spatial Data Quality (W. Shi, M.F. Goodchild, and P.F. Fisher, editors), Taylor & Francis, London, UK, p. 230-250.

Unwin D., "Geographical information systems and the problem of error and uncertainty", *Progress in Human Geography*, vol. 19, 1995, p. 549-558.

Bibliographie générale

Cette section contient l'ensemble des références consultées, ayant contribué à la présente thèse (Les références citées dans les chapitres de la thèse sont identifiées par un astérisque : *).

* Aalders H. J. G. L., "The registration of Quality in a GIS", *Proceedings of International Symposium on Spatial Data Quality*, Hong Kong, 18-20 July 1999, p. 23-32.

* Aalders H. J. G. L., "The Registration of Quality in a GIS". *Spatial Data Quality* (W. Shi, P. Fisher, and M. F. Goodchild, Eds), Taylor & Francis, p. 186-199, 2002.

* Aalders H. J. G. L., Morrison J., "Spatial Data Quality for GIS". *Geographic Information Research: Trans-Atlantic Perspectives* Eds), Taylor & Francis, London/Bristol, p. 463-475, 1998.

Aamodt A., Plaza E., "Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches", *AI Communications*, vol. 7, nº 1, 1994, p. 39-59.

* Agumya A., Hunter G. J., "Determining fitness for use of geographic information", *ITC Journal*, vol. 2, nº 1, 1997a, p. 109-113.

Agumya A., Hunter G. J., "Estimating Risk in GIS-Supported Decisions", *Proceedings of URISA*, Toronto, Canada, July 1997, p.

Agumya A., Hunter G. J., "Fitness for use: Reducing the Impact of Geographic Information Uncertainty", *Proceedings of URISA*, Charlotte, USA, 1998, p. 245-254.

* Agumya A., Hunter G. J., "Assessing "fitness for use" of geographic information: What risk are we prepared to accept in our decisions ?" *Spatial Accuracy Assessment, Land Information Uncertainty in Natural Resources* (K. Lowell, and A. Jaton, Eds), Quebec, p. 35-43, 1999a.

* Agumya A., Hunter G. J., "A Risk-Based Approach to Assessing the 'Fitness for Use' of Spatial Data", *URISA Journal*, vol. 11, nº 1, 1999b, p. 33-44.

Agumya A., Hunter G. J., "Translating Uncertainty in Geographical Data into Risk in Decisions", *Proceedings of 1st International Symposium on Spatial Data Quality*, Hong Kong, 18-20 July 1999, p. 574-584.

* Agumya A., Hunter G. J., "Responding to the consequences of uncertainty in geographical data", *International Journal of Geographical Information Science*, vol. 16, nº 5, 2002, p. 405-417.

Albaredes G., "A New Approach: User Oriented GIS", *Proceedings of EGIS '92*, Munich, p. 830-837.

Azouzi M., Merminod B., "Qualité des données spatiales", *Vermessung, Photogrammetrie, Kulturtechnik*, vol. 12, 1996, p. 645-649.

* Baddeley A., *Human Memory: Theory and Practice*, East Sussex, U.K., U.K. Psychology Press, 1997.

Bard, S., "Quality Assessment of Cartographic Generalisation", *Transactions in GIS*, vol.8, p. 63-81.

Bartsh-Spörl B. Lenz M., Hübner A., "Case-Based Reasoning - Survey and Future Directions", *Proceedings of XPS-99: Knowledge-Based Systems, Survey and Future Directions*, Würzburg, Germany, March 3-5, 1999, Springer, p. 67-89.

* Beard K., "Use error: the neglected error component", *Proceedings of AUTO-CARTO 9*, Baltimore, Maryland, March, 1989, p. 808-817.

* Beard K., "Representations of Data Quality". *Geographic Information Research: Bridging the Atlantic* (M. Craglia, and H. Couclelis, Eds), Taylor and Francis, p. 280-294, 1997.

Beard K., "Roles of Meta-Information in Uncertainty Management". *Mapping Ecological Uncertainty - Implications for Remote Sensing and GIS Applications* (C. T. Hunsaker, M. F. Goodchild, M. A. Friedl, and T. J. Case, Eds), Springer-Verlag, p. 363-378, 2001.

* Beard K., Buttenfield B., "Detecting and evaluating errors by graphical methods". *Geographical Information Systems* (P. A. Longley, M. F. Goodchild, D. J. Maguire, and D. W. Rhind, Eds), Wiley, p. 219-233, 1999.

* Beard K., Mackaness W., "Visual Access to Data Quality in Geographic Information Systems", *Cartographica*, vol. 30, nº 2-3, 1993, p. 37-45.

Beard K., Sharma V., "Multilevel and Graphical Views of Metadata", *Proceedings of IEEE Advances in Digital Libraries (ADL)*, Santa-Barbara, USA, p. 256-265, 1998.

* Bédard Y., *A Study of the Nature of Data Using a Communication-Based Conceptual Framework of Land Information Systems*, PhD Thesis, University of Maine, Orono, 1986.

* Bédard Y., "Uncertainties in Land Information Systems Databases", *Proceedings of Eighth International Symposium on Computer-Assisted Cartography*, Baltimore, Maryland, March 29th - April 3rd 1987, American Society for Photogrammetry and Remote Sensing and American Congress on Surveying and Mapping, p. 175-184.

Bédard Y., "Towards Collaborative Research Projects in Geomatics Applied to Health Surveillance", *Proceedings of Tri-Council Workshop/Networking Program*, Centre for Research in Geomatics, Laval University, Quebec City, October 2000.

Bédard Y. Devillers R., Gervais M., "Vers une gestion et communication dynamique des informations sur la qualité des données géospatiales", *Proceedings of Géomatique 2002*, Montréal, Canada, 30 Octobre 2002.

Bédard Y. Devillers R. Gervais M., Jeansoulin R., "Towards Multidimensional User Manuals for Geospatial Datasets: Legal issues and their Considerations into the design of a Technological Perspective", *Proceedings of 3rd International Symposium on Spatial Data Quality (ISSDQ'04)*, Bruck an der Leitha, Austria, April 15-17th 2004, p. 183-195.

* Bédard Y. Gosselin P. Rivest S. Proulx M.-J. Nadeau M. Lebel G., Gagnon M.-F., "Integrating GIS Components with Knowledge Discovery Technology for Environmental Health Decision Support", *International Journal of Medical Informatics*, vol. 70, nº 1, 2003, p. 79-94.

Bédard Y. Merrett T., Han J., "Fundamentals of Spatial Data Warehousing for Geographic Knowledge Discovery". *Geographic Data mining and Knowledge Discovery* (H. Miller, and J. Han, Eds), Taylor & Francis, 2001a.

Bédard Y. Proulx M.-J., Larrivée S., Qualité des données à référence spatiale, 2001b.

* Bédard Y., Vallière D., 1995. Qualité des données à référence spatiale dans un contexte gouvernemental, Rapport de recherche, Université Laval, Québec, Canada.

Bédard Y. Vallière D., Métivier R., "Nouvelle méthode d'évaluation de la qualité des données à référence spatiale", *Proceedings of 8e Conférence internationale sur la géomatique*, Ottawa, May 28-30th 1996.

Bernhardsen T., "Choosing a GIS". *Geographical Information Systems* (P. A. Longley, M. F. Goodchild, D. J. Maguire, and D. W. Rhind, Eds), John Wiley & Sons, Inc., p. 589-600, 1999.

* Berry B., "Approaches to regional analysis: a synthesis." *Annals of the Association of American Geographers*, vol. 54, 1964, p. 2-11.

* Berson A., Smith S. J., *Data Warehousing, Data Mining and OLAP (Data Warehousing / Data Management)*, New-York, McGraw-Hill, 1997.

* Bertin J., *Sémiologie graphique: les diagrammes, les réseaux, les cartes*, Paris, Mouton-Gauthier-Villars-Bordas, 1973.

* Blackmore M., "High or Low Resolution? Conflicts of Accuracy, Cost, Quality and Application in Computer Mapping", *Computers & Geosciences*, vol. 11, nº 2, 1985, p. 345-348.

Body M. Miquel M. Bédard Y., Tchounikine A., "Handling Evolutions in Multidimensional Structures", *Proceedings of 19th International Conference on Data Engineering (ICDE)*, Bangalore, India, 5-8 March 2003.

* Box G. E. P., "Science and statistics", *Journal of the American Statistical Association*, vol. 71, 1976, p. 791-799.

Brassel K. Bucher F. Stephan E.-M., Vckovski A., "Completeness". *Elements of spatial data quality* (S. C. Guptill, and J. L. Morrison, Eds), p. 81-108, 1995.

* Brodeur J. Bédard Y. Edwards G., Moulin B., "Revisiting the Concept of Geospatial Data Interoperability within the Scope of Human Communication Processes", *Transactions in GIS*, vol. 7, nº 2, 2003, p. 243-265.

Brodeur J., Massé F., "Standardization in Geomatics: in Canada and in ISO/TC 211", *Geomatica*, vol. 55, nº 1, 2001, p. 91-106.

Brown J. Heuvelink G. B. M., Refsgaard J. C., "Assessing and recording uncertainties about environmental data", *Proceedings of Third International Symposium on Spatial Data Quality (ISSDQ 04)*, Bruck an der Leitha, Austria, GeoInfo Series, p. 249-259, 2004.

* Burrough P. A., "Development of intelligent geographical information systems", *International Journal of Geographical Information Systems*, vol. 6, nº 1, 1992, p. 1-11.

Buttenfield B., "Spatial Uncertainty in Ecology". *Mapping Ecological Uncertainty - Implications for Remote Sensing and GIS Applications* (C. T. Hunsaker, M. F. Goodchild, M. A. Friedl, and T. J. Case, Eds), Springer-Verlag, p. 115-132, 2001.

* Buttenfield B., Beard K. M., "Graphical and Geographical components of Data Quality". *Visualization in Geographic Information Systems* (H. M. Hearnshaw, and D. J. Unwin, Eds), Wiley, p. 150-157, 1994.

* Buttenfield B. P., "Representing Data Quality", *Cartographica*, vol. 30, nº 2-3, 1993, p. 1-7.

* Buttenfield B. P., Beard K., "Visualizing the quality of spatial information", *Proceedings of AUTO-CARTO 10*, p. 423-427, 1991.

Caron P.-Y., *Étude du potentiel de OLAP pour supporter l'analyse spatio-temporelle*, Mémoire, Université Laval, Québec, 1998.

* CEN/TC-287, 1994/1995. WG 2, Data description: Quality. Working paper N. 15, August 1994. PT05, Draft Quality Model for Geographic Information, Working paper D3, January 1995.

Charnay L., *Dialogue et explication dans les systèmes à base de connaissances - ADex, un modèle informatique pour l'énonciation*, Thèse de doctorat, U. Orsay, Paris, 1999.

* Charron J., *Développement d'un processus de sélection des meilleures Sources de données cartographiques pour leur intégration à une base de données à référence spatiale*, Mémoire, Université Laval, Québec, 1995.

* Chrisman N. R., "The Role of Quality information in the Long Term Functioning of a Geographical Information System." *Proceedings of International Symposium on Automated Cartography (Auto Carto 6)*, Ottawa, Canada, p. 303-321.

* Chrisman N. R., "The error component in spatial data". *Geographic Information Systems: Principles and Applications* (D. J. Maguire, M. F. Goodchild, and D. W. Rhind, Eds), Wiley, London, p. 165-174, 1990.

Chrisman N. R., *Exploring Geographic Information Systems*, John Wiley & Sons, 1997.

* Chrisman N. R., "Speaking Truth to Power: An Agenda for Change". *Spatial Accuracy Assessment, Land Information Uncertainty in Natural Resources* (K. Lowell, and A. Jaton, Eds), Quebec, p. 27-31, 1998.

Clarke D. G., Clark D. M., "Lineage". *Elements of spatial data quality* (S. C. Guptill, and J. L. Morrison, Eds), p. 13-30, 1995.

Clarke K. C., Teague P. L., "Representation of Cartographic Uncertainty Using Virtual Environments", *Proceedings of 4th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, Amsterdam, Pays-Bas, Juillet 2000, p. 109-116.

* Codd E. F., 1993. Providing OLAP (On-line Analytical Processing) to User-Analysts: An IT Mandate, E. F. Codd and Associates.

* CTG, 2000. Insider's Guide to Using Information in Government - The devil is in the data, Center for Technology in Government.

* Curry M. R., *Digital places: Living with Geographic Information Technologies*, London & New-York, Routeledge, 1998.

Dassonville L., "Quality Management, data quality and users, metadata for geographical information", *Proceedings of International Symposium on Spatial Data Quality*, Hong Kong, 18-20 July 1999, p. 133-143.

* Dassonville L. Vauglin F. Jakobsson A., Luzet C., "Quality Management, Data Quality and Users, Metadata for Geographical Information". *Spatial Data Quality* (W. Shi, P. Fisher, and M. F. Goodchild, Eds), Taylor & Francis, p. 202-215, 2002.

* David B., Fasquel P., 1997. Bulletin d'information de l'IGN - Qualité d'une base de données géographique: concepts et terminologie, N. 67, IGN France.

Davis T. J., Keller P., "Modelling and Visualizing Multiple Spatial Uncertainties", *Computer and Geosciences*, vol. 23, nº 4, 1997, p. 397-408.

* De Bruin S. Bregt A., Van de Ven M., "Assessing fitness for use: the expected value of spatial data sets", *International Journal of Geographical Information Science*, vol. 15, nº 5, 2001, p. 457-471.

De Groeve T., *L'incertitude spatiale dans la cartographie forestière*, Ph.D. Thesis, Université Laval, Québec, 1999.

Drecki I., "Visualisation of Uncertainty in Geographic Data", *Proceedings of International Symposium on Spatial Data Quality*, Hong Kong, 18-20 July 1999, p. 260-271.

* Drecki I., "Visualisation of Uncertainty in Geographic Data". *Spatial Data Quality* (W. Shi, P. F. Fisher, and M. F. Goodchild, Eds), Taylor & Francis, p. 140-159, 2002.

Drummond J., "Positional accuracy". *Elements of spatial data quality* (S. C. Guptill, and J. L. Morrison, Eds), p. 31-58, 1995.

Duckham M., "Implementing an object-oriented error-sensitive GIS", *Proceedings of Spatial accuracy assessment: land information uncertainty in natural resources*, Québec, Canada, p. 209-215, 1998.

Duckham M., "A user-oriented perspective of error-sensitive GIS development", *Transactions in GIS*, vol. 6, nº 2, 2002, p. 179-194.

Duckham M. Drummond J., Forrest D., "Spatial data quality capture through inductive learning", *Spatial Cognition and Computation*, vol. 2, nº 4, 2000, p. 261-282.

* Duckham M., McCreadie J., "An intelligent, distributed, error-aware OOGIS", *Proceedings of 1st International Symposium on Spatial Data Quality*, Hong Kong, 18-20 July 1999, p. 496-506.

* Duckham M., McCreadie J. E., "Error-aware GIS Development". *Spatial Data Quality* (W. Shi, P. F. Fisher, and M. F. Goodchild, Eds), Taylor & Francis, London, p. 63-75, 2002.

* Eco U., "De l'impossibilité d'établir une carte de l'empire à l'échelle de 1/1". *Pastiches et Postiches* (U. Eco, Eds), Éditions 10/18, p. 183, 2000.

Edwards G., Fortin M.-J., "A Cognitive View of Spatial Uncertainty". *Mapping Ecological Uncertainty - Implications for Remote Sensing and GIS Applications* (C. T. Hunsaker, M. F. Goodchild, M. A. Friedl, and T. J. Case, Eds), Springer-Verlag, p. 133-157, 2001.

Elmes G. A., Cai G., "Data Quality Issues in User Interface Design for a Knowledge-Based Decision Support System", *Proceedings of Fith International Symposium on Spatial Data Handling*, Charleston, USA, p. 303-312.

* Elshaw Thrall S., Thrall G. I., "Desktop GIS software". *Geographical Information Systems* (P. A. Longley, M. F. Goodchild, D. J. Maguire, and D. W. Rhind, Eds), John Wiley & Sons, New-York, p. 331-345, 1999.

* Epstein E. F. Hunter G. J., Agumya A., "Liability insurance and the use of geographical information", *International Journal of Geographical Information Science*, vol. 12, nº 3, 1998, p. 203-214.

Faïz S. Abbassi K., Boursier P., "Applying Data Mining Techniques to Generate Quality Information from Geographical Databases". *Data Quality in Geographic Information - From Error to Uncertainty* (M. F. Goodchild, and R. Jeansoulin, Eds), Editions Hermes, p. 192, 1998.

Faïz S., Zghal H. B., "Managing Quality by using OLAP Techniques and Data Warehouses", *Proceedings of Accuracy 2000*, Amsterdam, July 2000, p. 203-206.

* Faïz S. O., *Modélisation, exploitation et visualisation de l'information qualité dans les bases de données géographique*, Ph.D. thesis, Université Paris-Sud, Paris, 1996.

* Faïz S. O., *Systèmes d'Informations Géographiques: Information Qualité et Data Mining*, Tunis, Editions C.L.E, 1999.

* Fernandez A., *Les nouveaux tableaux de bord des décideurs*, Paris, Éditions d'organisation, 2000.

* FGDC, 2000. Content Standard for Digital Geospatial Metadata Workbookversion 2.

Fischhoff B. Lichtenstein S. Slovic P. Derby S. L., Keeney R. L., *Acceptable risk*, Cambridge (UK), Cambridge University Press, 1981.

* Fisher G., "User Modeling in Human-Computer Interaction", *User Modeling and User-Adapted Interaction*, vol. 11, 2001, p. 65-86.

* Fisher P., "Animation and sound for the visualization of uncertain spatial information". *Visualization in Geographic Information Systems* (H. M. Hearnshaw, and D. J. Unwin, Eds), Wiley, p. 181-185, 1994a.

* Fisher P., "Visualising the uncertainty of soil maps by animation", *Cartographica*, vol. 30, 1994b, p. 20-27.

* Fisher P. F., "Models of uncertainty in spatial data". *Geographical Information Systems* (P. A. Longley, M. F. Goodchild, D. J. Maguire, and D. W. Rhind, Eds), John Wiley & Sons, New-York, p. 191-205, 1999.

* Frank A. U., "Metamodels for Data Quality Description". *Data Quality in Geographic Information - From Error to Uncertainty* (M. F. Goodchild, and R. Jeansoulin, Eds), Editions Hermes, p. 192, 1998.

* Frank A. U. Grum E., Vasseur B., "How to select the Best Dataset for a Task?" *International Journal of Geographical Information Science*, vol., Submitted.

* Gan E., Shi W., "Error Metadata Management System". *Spatial Data Quality* (W. Shi, P. F. Fisher, and M. F. Goodchild, Eds), Taylor Francis, London and New York, p. 336, 2002.

* Gervais M., *Pertinence d'un manuel d'instructions au sein d'une stratégie de gestion du risque juridique découlant de la fourniture de données géographiques numériques*, Ph.D. thesis, Université Laval, Québec, 2004.

* Gervais M. Devillers R. Bédard Y., Jeansoulin R., "GI Quality and Decision making : toward a contextual user manual", *Proceedings of GeoInformation Fusion and Revision Workshop*, Quebec city, Canada, April 9-12, 2001.

* Goglin J.-F., *Le datawarehouse pivot de la relation client*, Paris, France, Hermès Sciences, 2001.

Goodchild M. F., "Attribute accuracy". *Elements of spatial data quality* (S. C. Guptill, and J. L. Morrison, Eds), p. 59-79, 1995a.

* Goodchild M. F., "Sharing Imperfect Data". *Sharing Geographic Information* (H. J. Onsrud, and G. Rushton, Eds), Rutgers University Press, New Brunswick, NJ, p. 413-425, 1995b.

Goodchild M. F., "Measurement-based GIS". *Spatial Data Quality* (W. Shi, P. F. Fisher, and M. F. Goodchild, Eds), Taylor & Francis, London, p. 5-17, 2002.

* Goodchild M. F. Buttenfield B., Wood J., "Introduction to visualizing data validity". *Visualization in Geographic Information Systems* (H. M. Hearnshaw, and D. J. Unwin, Eds), Wiley, p. 141-149, 1994a.

Goodchild M. F. Chih-Chang L., Leung Y., "Visualizing fuzzy maps". *Visualization in Geographical Information Systemss* (H. M. Hearnshaw, and D. Unwin, Eds), Wiley, Chichester, p. 158-167, 1994b.

* Goodchild M. F., Kemp K. K., 1990. NCGIA Core Curriculum in GIS, National Center for Geographic Information and Analysis, University of California, Santa Barbara CA.

* Gottsegen J. Montello D., Goodchild M. F., "A Comprehensive Model of Uncertainty in Spatial Data", *Proceedings of Spatial Accuracy Assessment: Land Information Uncertainty in Natural Resources*, Québec, Canada, Ann Arbor Press, p. 175-182, 1998.

Gruber T. R., "A Translation Approach to Portable Ontology Specifications", *Knowledge Acquisition*, vol. 5, nº 2, 1993, p. 199-220.

* Grum E., Vasseur B., "How to select the best dataset for a task?" *Proceedings of 3rd International Symposium on Spatial Data Quality (ISSDQ'04)*, Bruck an der Leitha, Austria, April 15-17th, GeoInfo Series, p. 197-206, 2004.

Guptill S., "Building a Geospatial Data Framework - Finding the Best Available Data". *Data Quality in Geographic Information - From Error to Uncertainty* (M. F. Goodchild, and R. Jeansoulin, Eds), Editions Hermes, p. 192, 1998.

Guptill S. C., "Temporal information". *Elements of spatial data quality* (S. C. Guptill, and J. L. Morrison, Eds), p. 153-166, 1995.

* Guptill S. C., "Metadata and data catalogues". *Geographical Information Systems* (P. A. Longley, M. F. Goodchild, D. J. Maguire, and D. W. Rhind, Eds), John Wiley & Sons, Inc., p. 677-692, 1999.

* Guptill S. C., Morrison J. L., *Elements of spatial data quality*, New York, Elsevier Science, 1995.

* Harvey F., "Quality Needs More Than Standards". *Data Quality in Geographic Information - From Error to Uncertainty* (M. F. Goodchild, and R. Jeansoulin, Eds), Editions Hermes, p. 192, 1998.

Hennings V., Boess J., "User-oriented Concepts to Assess the Accuracy of Nationwide Land Quality Maps", *Proceedings of 4th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, Amsterdam, Pays-Bas, p. 301-304, 2000.

* Heuvelink G. B. M., Lemmens M. J. P. M., *4th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, Amsterdam, The Nederland, 2000.

* Holmwood T. S., "Data Quality: Defining an achievable standard", *Proceedings of GITA Annual conference*, 2000.

Holt A., Benwell G. L., "Using Spatial Similarity for Exploratory Spatial Data Analysis: Some Directions", *Proceedings of GeoComputation '97 and SIRC '97*, Otago, New Zealand, 26-29 August 1997, p. 15-24.

Holt A., Benwell G. L., "Applying case-based reasoning techniques in GIS", *International Journal of Geographical Information Science*, vol. 13, nº 1, 1999, p. 9-25.

Hoxmeier J. A., "Typology of database quality factors", *Software Quality Journal*, vol. 7, 1998, p. 179-193.

Hunsaker C. T. Goodchild M. F. Friedl M. A., Case T. J., ed., 2001. *Mapping Ecological Uncertainty - Implications for Remote Sensing and GIS Applications*, Springer-Verlag, 402 p.

Hunter A., *Uncertainty in Information Systems*, 1996.

* Hunter G. J., "Managing uncertainty in GIS". *Geographical Information Systems* (P. A. Longley, M. F. Goodchild, D. J. Maguire, and D. W. Rhind, Eds), John Wiley & Sons, Inc., p. 633-641, 1999a.

Hunter G. J., "New Tools For Handling Spatial Data Quality: moving from Academic Concepts to Practical Reality", *URISA Journal*, vol. 11, nº 2, 1999b.

* Hunter G. J., "Spatial Data Quality Revisited", *Proceedings of GeoInfo 2001*, Rio de Janeiro, Brazil, 4-5th October, p. 1-7.

Hunter G. J., "Understanding Semantics and Ontologies: They're Quite Simple Really - If You Know What I Mean", *Transactions in GIS*, vol. 6, nº 2, 2002, p. 83-87.

* Hunter G. J., Lowell K., *5th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, Melbourne, Australia, 2002.

* Hunter G. J., Masters E., "What's Wrong with Data Quality Information?" *Proceedings of GIScience 2000*, Savannah, USA, p. 201-203, 2000.

* Hunter G. J., Reinke K. J., "Adapting Spatial Databases to Reduce Information Misuse Through Illogical Operations", *Proceedings of 4th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences (Accuracy 2000)*, Amsterdam, July 2000, p. 313-319.

Hunter G. J. Wachowicz M., Bregt A. K., "Understanding Spatial Data Usability", *Data Science Journal*, vol. 2, 2003, p. 79-89.

* ISO 8402, 1994. Quality management and quality assurance - Vocabulary, International Organization for Standardization (ISO).

* ISO-TC/211, 2002. Geographic Information - Quality principles19113.

* ISO-TC/211, 2003a. Geographic Information - Metadata19115.

* ISO-TC/211, 2003b. Geographic Information - Quality evaluation procedures19114.

Jakobsson A., "Quality Evaluation of Topographic Datasets - Experiences in European National Mapping Agencies", *Proceedings of International Symposium on Spatial Data Quality*, Hong Kong, 18-20 July 1999, p. 154-164.

Jakobsson A., Vauglin F., "Status of Data Quality in European National Mapping Agencies", *CFC*, vol., nº 169-170, 2001a, p. 21-26.

Jakobsson A., Vauglin F., "Status of Data Quality in European National Mapping Agencies", *Bulletin de la Commission Française de Cartographie (CFC)*, vol. 169-170, 2001b, p. 21-26.

Jarke M., Vassiliou Y., "Data Warehouse Quality: A Review of the DWQ Project", *Proceedings of 2nd Conference on Information Quality*, Cambridge, USA, p. 299-313, 1997.

Jeansoulin R., Papini O., "Révision et systèmes d'informations géographiques". *Le Temps, l'Espace, l'Evolutif, dans les sciences du traitement de l'information* (Cepadues, Eds), Toulouse, p. 293-304, 2000.

* Juran J. M. Gryna F. M. J., Bingham R. S., *Quality Control Handbook*, New-York, McGraw-Hill, 1974.

* Kahn B. K., Strong D. M., "Product and Service Performance Model for Information Quality: An Update." *Proceedings of Conference on Information Quality*, Cambridge, MA: Massachusetts Institute of Technology.

Kainz W., "Logical consistency". *Elements of spatial data quality* (S. C. Guptill, and J. L. Morrison, Eds), p. 109-137, 1995.

* Kaplan R., Norton D., "The balanced scorecard: Measures that Drive Performance", *Harvard Business Review*, vol. 70, nº 1, 1992, p. 71-79.

Keller S. F., "On the Use of Case-Based Reasoning in Generalization", *Proceedings of Spatial Data Handling 6*, Edinburgh, Scotland, UK, 5th-9th September 1994, p. 1118-1132.

* Klein G., *Sources of Power - How people make decisions*, Cambridge, Massachusetts, MIT Press, 1999.

* Kobsa A., "Generic User Modeling Systems", *User Modeling and User-Adapted Interaction*, vol. 11, 2001, p. 49-63.

* Krek A., Frank A. U., "Optimization of Quality of Geoinformation Products", *Proceedings of Proceedings of 11th Annual Colloquium of the Spatial Information Research Centre, SIRC'99*, Dunedin, New Zealand, 13-15 December, 1999, Dept. of Information Science, University of Otago, p. 151-159.

Lanter D., "A Three-part Approach to Geographic Data Quality Assurance". *Data Quality in Geographic Information - From Error to Uncertainty* (M. F. Goodchild, and R. Jeansoulin, Eds), Editions Hermes, p. 192, 1998.

Larsen P. L., "Learning to Speak Metadata", *GIS Europe*, vol. July, 1996, p. 20-22.

Lee Y. C., Chan H. C. E., "Spatial Metadata and its Management", *Geomatica*, vol. 54, nº 4, 2000, p. 451-462.

* Leitner M., Buttenfield B. P., "Guidelines for the Display of Attribute Certainty", *Cartography and Geographic Information Science*, vol. 27, nº 1, 2000, p. 3-14.

Lemon O., Pratt I., "Logics for geographic information", *Journal of Geographical Systems*, vol. 1, 1999, p. 75-90.

* Létourneau F. Bédard Y., Moulin B., "Perspectives d'utilisation du concept d'entrepôt de données pour les géorépertoires dans internet", *Geomatica*, vol. 52, nº2, 1998, p. 145-163.

Lilburne L., Benwell G., "The Scale Matcher: Determining Scale Compatibility of Environmental Data and Models", *Proceedings of 4th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, Amsterdam, Pays-Bas, Juillet 2000, p. 417-424.

* Longley P. A. Goodchild M. F. Maguire D. J., Rhind D. W., ed., 1999. *Geographical Information Systems*, John Wiley & Sons

* Longley P. A. Goodchild M. F. Maguire D. J., Rhind D. W., ed., 2001. *Geographical Information Systems and Science*, John Wiley & Sons, 454 p.

Loriette-Rougegrez S., "Raisonnement à partir de cas pour les évolutions spatiotemporelles de processus", *Revue internationale de géomatique*, vol. 8, nº 1-2, 1998, p. 207-227.

* Lowell K., "Why aren't we making better use of uncertainty information in decision-making?" *Proceedings of 6th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, Portland, Maine, USA, 2004.

* Lowell K., Jaton A., *3rd International on Spatial Accuracy Assessment, Land Information Uncertainty in Natural Resources*, Ann Arbor Press, Quebec, Canada, 1999.

Luger G., Stubblefield W. A., *Artificial Intelligence - Structures and Strategies for Complex Problem Solving*, Addison Wesley, 1999.

* Mac Eachren A. M., "Visualizing uncertain information", *Cartographic Perspectives*, vol. 13, 1992, p. 10-19.

MacEachren A. Bishop I. Dykes J. Dorling D., Gatrell A., "Introduction to Advances in Visualizing Spatial Data". *Visualization in Geographic Information Systems* (H.M. Hearnshaw and D.J. Unwin, Eds), Wiley, p. 51-59, 1994.

MacEachren A., Kraak M.-J., "Exploratory Cartographic Visualization: Advancing the Agenda", *Computer and Geosciences*, vol. 23, nº 4, 1997, p. 335-343.

Malczewski J., *GIS and Multicriteria Decision Analysis*, New York, Wiley, 1999.

* Manche Y., *Analyse spatiale et mise en place de systèmes d'information pour l'évaluation de la vulnérabilité des territoires de montagne face aux risques naturels*, Thèse de doctorat, Université Joseph Fourier, Grenoble, 2000.

* Martinet B., Marti Y.-M., *L'intelligence économique*, Éditions d'Organisation, 2001.

* McGranaghan M., "A cartographic View of Spatial Data Quality", *Cartographica*, vol. 30, nº 2-3, 1993, p. 8-19.

Medyckyj-Scott D., Hearnshaw H.M., ed., 1993. *Human Factors in Geographical Information Systems*, Belhaven Press, 266 p.

Meng L., "Scroll the space and drill-down the information", *Proceedings of 20th International Cartographic Conference*, Beijing, China, 6-10 août 2001, p. 2436-2443.

Mihaila G.A. Rashid L., Vidal M.E., "Querying "Quality of Data" Metadata", *Proceedings of Third IEEE META-DATA Conference*, Maryland, USA, avril 1999.

* Miller G.A., "The Magical Number Seven, plus or minus two: Some limits on our capacity for processing information", *The Psychological Review*, vol. 63, 1956, p. 81-97.

* Miller H.J., Han J., *Geographic Data Mining and Knowledge Discovery*, Taylor & Francis, 2001.

* Mintzberg H., *The structuring of organisations*, Prentice-Hall, 1979.

* Monmonier M., "A Case Study in the Misuse of GIS: Siting a Low-Level Radioactive Waste Disposal Facility in New-York State", *Proceedings of Conference on Law and Information Policy for Spatial Databases*, Tempe (AZ) USA, p. 293-303, 1994.

* Morrison J. L., "A theoretical framework for cartographic generalisation with the emphasis on the process of symbolisation", *International Yearbook of Cartography*, vol. 14, p. 115-127, 1974.

* Morrison J. L., "Spatial data quality". *Elements of spatial data quality* (S.C. Guptill, and J.L. Morrison, Eds), Elsevier Science inc., New York, 1995.

* Mowrer H. T., "Accuracy (Re)assurance: Selling Uncertainty Assessment to the Uncertain". *Spatial Accuracy Assessment, Land Information Uncertainty in Natural Resources* (K. Lowell, and A. Jaton, Eds), Quebec, p. 3-10, 1999.

Navratil G., "How Laws affect Data Quality", *Proceedings of Third International Symposium on Spatial Data Quality*, Bruck an der Leitha, Austria, GeoInfo Series, p. 37-47, 2004.

* Newell A., *Unified theories of cognition*, Cambridge, Harvard University Press, 1990.

Obermeyer N. J., "Measuring the benefits and costs of GIS". *Geographical Information Systems* (P.A. Longley, M.F. Goodchild, D.J. Maguire and D.W. Rhind, Eds), John Wiley & Sons, Inc., p. 601-610, 1999.

* Office québécois de la langue française, 2004. www.olf.gouv.qc.ca

Onsrud H. J., "Liability in the use of GIS and geographical datasets". *Geographical Information Systems* (P.A. Longley, M.F. Goodchild, D.J. Maguire and D.W. Rhind, Eds), John Wiley & Sons, Inc., p. 643-652, 1999.

Pang A., "Visualizing Uncertainty in Geo-spatial Data", *Proceedings of Workshop on the Intersections between Geospatial Information and Information Technology for the National Academies committee of the Computer Science and Telecommunications Board*, Arlington, USA, p. 1-14, 2001.

* Paradis J., Beard K., "Visualization of Spatial Data Quality for the Decision Maker: A Data Quality Filter", *URISA Journal*, vol. 6, nº 2, 1994, p. 25-34.

Peterson L. R., Peterson M. J., "Short-Term Retention of Individual Verbal Items", *Journal of Experimental Psychology*, vol. 58, nº 3, 1959, p. 193-198.

Peuquet D., "It's about time: A conceptual framework for the representation of temporal dynamics in geographic information systems", *Annals of the Association of American Geographers*, vol. 84, nº 3, 1994, p. 441-461.

* Plan Canada, 1999. Sustainable community indicators program Vol 39 (5).

Platon, *Les lois*.

Plewe B., "The Nature of Uncertainty in Historical Geographic Information", *Transactions in GIS*, vol. 6, nº 4, 2002, p. 431-456.

Pontikakis E., Frank A., "Basic Spatial Data According to User's Needs Aspects of Data Quality", *Proceedings of Third International Symposium on Spatial Data Quality*, Bruck an der Leitha, Austria, GeoInfo Series, p. 13-21, 2004.

* Proulx M. J., Bédard Y., "Le géorépertoire, un outil de gestion cartographique", *Arpenteur-Géomètre, Revue de l'Ordre des Arpenteurs-Géomètres du Québec*, vol. 21, nº 5, 1995, p. 21-24.

* Proulx M. J. Bédard Y. Létourneau F., Martel C., "Catalogage des données spatiales sur le world wide web: concepts, analyses des sites et présentation du géorépertoire personnalisable GEOREP", *Revue Internationale de Géomatique*, vol. 7, nº 1, 1997, p. 7-32.

* Qiu J., Hunter G. J., "Managing Data Quality Information", *Proceedings of International Symposium on Spatial Data Quality*, Hong Kong, 18-20 juillet 1999, p. 384-395.

Qiu J., Hunter G. J., "Towards Dynamic Updating of Data Quality Information", *Proceedings of Accuracy 2000*, Amsterdam, juillet 2000, p. 529-536.

* Qiu J., Hunter G. J., "A GIS with the Capacity for Managing Data Quality Information". *Spatial Data Quality* (W. Shi, M.F. Goodchild, and P.F. Fisher, Eds), Taylor & Francis, London, p. 230-250, 2002.

* Rafanelli M., *Multidimensional Databases: Problems and Solutions*, Hershey, USA, Idea Group Publishing, 2003.

Reinke K. J., Hunter G. J., "Communicating Quality in Spatial Information: Notification - the First Step", *Proceedings of International Symposium on Spatial Data Quality*, Hong Kong, 18-20 juillet 1999, p. 66-75.

* Reinke K. J., Hunter G. J., "A Theory for Communicating Uncertainty in Spatial Databases". *Spatial Data Quality* (W. Shi, P.F. Fisher, and M.F. Goodchild, Eds), Taylor & Francis, London, p. 77-101, 2002.

* REV!GIS, 2001. Uncertain Knowledge Maintenance and Revision in Geographic Information Systems, http://www.lsis.org/REVIGIS/.

* Rivest S. Bédard Y., Marchand P., "Towards Better Support for Spatial Decision Making: Defining the Characteristics of Spatial On-Line Analytical Processing (SOLAP)", *Geomatica*, vol. 55, nº 4, 2001, p. 539-555.

Roche V. Batton-Hubert M., Dechomets R., "Ambiguity and uncertainty in GIS design", *Proceedings of 4th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, Amsterdam, Pays-Bas, p. 549-551, 2000.

Salgé F., "Semantic accuracy". *Elements of spatial data quality* (S.C. Guptill and J.L. Morrison, Eds), p. 139-151, 1995.

Salgé F., "National and international standards". *Geographical Information Systems, Principles and Applications* (P.A. Longley, M.F. Goodchild, D.J. Maguire and D.W. Rhind, Eds), John Wiley & Sons, Inc., p. 693-706, 1999.

* Schramm W., "How Communication Works". *Communication: Concepts and Processes* (J.A. DeVito, Ed), Prentice-Hall, New Jersey, p. 12-21, 1971.

Schreiber G. Akkermans H. Anjewierden A. de Hoog R. Shadbolt N. Van de Velde W., Wielinga B., *Knowledge Engineering and Management - The CommonKADS Methodology*, Cambridge, Massachusetts, MIT Press, 2000.

* Shannon C. E., "A Mathematical Theory of Communication", *The Bell System Technical Journal*, vol. 27, 1948, p. 379-423.

* Simon H. A., "A Behavioral Model of Rational Choice?" *Quarterly Journal of Economics*, vol., nº 69, 1955, p. 99-118.

* Sinton D. F., "The inherent structure of information as a constraint in analysis". *Harvard papers on Geographic Information Systems* (G. Dutton, Ed), Addison-Wesley, Reading, USA, 1978.

* Smithson M., *Ignorance and Uncertainty: Emerging Paradigms*, New York, Springer Verlag, 1989.

Storey V. C., Wang R. Y., "Modeling Quality Requirements in Conceptual Database Design", *Proceedings of Third Conference on Information Quality*, Cambridge, USA, p. 64-87, 1998.

Sui D. Z., Goodchild M. F., "GIS as a Media?" *International Journal of Geographical Information Science*, vol. 15, nº 5, 2001, p. 387-390.

Swartout W. R., Moore J. D., "Explanation in Second Generation Expert Systems". *Second Generation Expert Systems* (J.-M. David, J.-P. Krivine and R. Simmons, Eds), Springer-Verlag, Berlin, New York, p. 543-585, 1993.

* Tastan H., Altan M. O., "Spatial Data Quality", *Proceedings of Third Turkish-German Joint Geodetic Days*, Istanbul, Turquie, 1-4 juin 1999, p. 15-30.

* Taylor J. R., *An introduction to error analysis: the study of uncertainties in physical measurements*, Oxford, University Science Books, 1982.

Thomsen E., *OLAP Solutions: Building Multidimensional Information Systems*, Wiley, 2002.

Thrill J.-C., ed., 1999. *Spatial Multicriteria Decision Making and Analysis*, Ashgate, 377 p.

* Timpf S. Raubal M., Kuhn W., "Experiences with Metadata", *Proceedings of Symposium on Spatial Data Handling, SDH'96, Advances in GIS Research II*, Delft, The Netherlands, 12-16 août 1996, IGU, p. 12B.31 - 12B.43.

Tsou M.-H., Buttenfield B. P., "An Agent-based, Global User Interface Distributed Geographic Information Services", *Proceedings of 8th International Symposium on Spatial Data Handling*, Vancouver, Canada, July 11-15th 1998, p. 603-612.

* Unwin D., "Geographical information systems and the problem of error and uncertainty", *Progress in Human Geography*, vol. 19, 1995, p. 549-558.

* Vasseur B., Devillers R., Jeansoulin R., "Ontological approach of the fitness of geospatial datasets", *Proceedings of 6th Agile Conference on Geographic Information Science*, Lyon, France, 24-26th April 2003, p. 497-504.

Vassiliadis P., Bouzeghoub M., Quix C., "Towards Quality-Oriented Data Warehouse Usage and Evolution", *Information Systems*, vol. 25, nº 2, 2000, p. 89-115.

Vauglin F., "A Practical Study on Precision and Resolution in Vector Geographical Databases", *Spatial Data Quality* (W. Shi, M.F. Goodchild, and P.F. Fisher, Eds), Taylor & Francis, London, p. 127-139, 2002.

* Veregin H., "Data quality parameters". *Geographical Information Systems* (P. A. Longley, M. F. Goodchild, D. J. Maguire, and D. W. Rhind, Eds), John Wiley & Sons, Inc., p. 177-189, 1999.

Veregin H., Hargitai P., "An evaluation matrix for geographical data quality". *Elements of spatial data quality*Eds), p. 167-188, 1995.

* von Schirnding Y. E., "Health-and-environment indicators in the context of sustainable development", *Proceedings of Consensus Conference on Environmental Health Surveillance: Agreeing on basic set of indicators and their future use*, Quebec city, Canada, October 10-12 2000.

* Voyer P., *Tableaux de bord de gestion et indicateurs de performance*, Presse de l'Université du Québec, 2000.

Wachowicz M., Hunter G. J., "Spatial Data Usability", *Data Science Journal*, vol. 2, 2003, p. 75-78.

* Wang R. Y., Strong D. M., "Beyond Accuracy: What Data Quality Means to Data Consumers", *Journal of Management Information Systems*, vol. 12, nº 4, 1996, p. 5-34.

Watson I., "An Introduction to Case-Based Reasoning", *Proceedings of Progress in Case-Based Reasoning*, Salford, UK, January 12, 1995, Springer, p. 3-16.

Weber R. Aha D. W., Becerra-Fernandez I., "Intelligent lessons learned systems", *Expert Systems with Applications*, vol. 17, 2001, p. 17-34.

Weber R. Aha D. W. Branting L. K. Lucas J. R., Fernandez I.-B., "Active Case-Based Reasoning for Lessons Delivery Systems", *Proceedings of AAAI-2000 Workshop on Intelligent Lessons Learned*, Menlo Park, AAAI Press, 2000.

* Willett G., *La communication modélisée - Une introduction aux concepts, aux modèles et aux théories*, Ottawa, 1992.

* Windholz T. K., *Strategies for Handling Spatial Uncertainty due to Discretization*, Ph.D. Thesis, University of Maine, Orono, 2001.

ANNEXE

Annexe 1

Table 2 : Liste des abréviations utilisées dans la thèse

Acronyme (français ou *anglais*)	**Signification**
BI	*Business Intelligence*
BNDT	Base Nationale de Données Topographiques
CEN	Comité Européen de Normalisation
CGSB/COG	*Canadian General Standard Board / Committee on Geomatics*
CIT-S	Centre d'Information Topographique de Sherbrooke
CRG	Centre de Recherche en Géomatique
CTG	*Center for Technology in Government*
DBMS	*DataBase Management System*
EIS	*Executive Information System*
ESRI	*Environmental Systems Research Institute*
FGDC	*Federal Geographic Data Committee*
GEOIDE (*GEOIDE)*	*Geomatics for Informed Decisions*
GIS	*Geographical Information System*
GPS	*Global Positioning System*
HOLAP	*Hybrid OLAP*
ICA	*International Cartographic Association*
IDG	Infrastructure de données géospatiales
IEEE	*Institute of Electrical and Electronics Engineers*
IGN	Institut Géographique National
ISO-TC	*International Organization for Standardization*
IST	*Information Society Technologies*
LBS	*Location-Based Services*
MDX	*Multidimensional Expressions Language*
MOLAP	*Multidimensional OLAP*

MUM (*MUM*)	Manuel à l'Usager Multidimensionel / *Multidimensional User Manual*
NCDCDS	*National Committee For Digital Cartographic Data Standards*
NCGIA	*National Centre for Geographic Information & Analysis*
NTBD	*National Topographic Database*
OGC	*Open Geospatial Consortium*
OLAP	*On-Line Analytical Processing*
OLTP	*On-Line Transactional Processing*
QIMM	*Quality Information Management Model*
ROLAP	*Relational OLAP*
RPD	*Recognition-Primed Decision*
SDTS	*Spatial Data Transfer Standard*
STM	*Short-Term Memory*
SIG	Système d'Information Géographique
SMMS	*Spatial Metadata Management System*
SOLAP (*SOLAP*)	OLAP Spatial / *Spatial OLAP*
SQL	*Structured Query Language*
XML	*Extensible Markup Language*